Patterns and processes of evolution at silent sites in mammalian genes

Chamary, Jean-Vincent

Award date:
2005

Awarding institution:
University of Bath

Patterns and processes of evolution at silent sites in mammalian genes

Jean-Vincent Chamary

A thesis submitted for the degree of Doctor of Philosophy
University of Bath
Department of Biology and Biochemistry
September 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Contents

Acknowledgements

Collaborator contribution

Unless otherwise stated, all the work presented in this thesis is my own. Martin Lercher did the majority of the analyses in Chapter 2. I provided tests and text relating to intron evolution and wrote several paragraphs of the introduction. Laurence Hurst identified putative regulatory regions within introns for Chapter 3. Joanna Parmley contributed equally to Chapter 6, in which tests for the human-mouse comparison were done in parallel. Jo also performed the simulations and analysed amino acid usage, while I examined human-chimp data. For Chapter 7, Laurence co-wrote the manuscript and constructed Figure 3, while Jo compiled Table 1.

Appreciation and dedication

I would like to thank several people for the help that they provided during the course of my Ph.D. Thanks to the researchers that I met at conferences who spared the time to comment upon my projects and discuss their work, science and life in general. I am simultaneously grateful and apologise to Amanda Harper, Richard Hooley and Richard ffrench-Constant for putting-up with my last-minute requests. Lastly, I am indebted to Laurence for his guidance, not only in how to ask the right questions, but also for showing me how to present the answers in an interesting way.

My work was funded by a studentship from the Biotechnology and Biological Sciences Research Council.

This thesis is dedicated to my parents, who will no doubt like the idea of having another doctor in the house.

Abbreviations

A	adenine
AT	bases that are either adenine or thymine
bp	base pairs
C	cytosine
cDNA	complementary DNA
ΔG	free energy
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
EST	expressed sequence tag
G	guanine
GC	bases that are either guanine or cytosine
ISE	intronic splicing enhancer
K_4	the number of substitutions at 4-fold degenerate sites per 4-fold site
K_a	the number of non-synonymous substitutions per non-synonymous site
Kb	kilobases
K_i	the number of substitutions per intronic site
K_{indel}	the number of insertions or deletions per base pair of an alignment
K_s	the number of synonymous substitutions per synonymous site
Mb	megabases
mRNA	messenger RNA
ncRNA	non-coding RNA
nt	nucleotides
ρ	Spearman rank correlation coefficient (called $r_{Spearman}$ in Chapter 2)
SAGE	serial analysis of gene expression
SINE	short interspersed element
SNP	single nucleotide polymorphism
T	thymine
TE	transposable element
TF	transcription factor
tRNA	transfer RNA
U	uracil
UTR	untranslated region

Summary

Within genes, mutations that do not change proteins are usually considered 'silent' because it is assumed that they have no affect on an organism's phenotype. Evolution at silent sites is expected to be selectively neutral, so that the probability that a new mutation will become fixed in a population is dictated by chance. This is especially important in mammals, where small effective population sizes should reduce the ability of natural selection to influence the fate of mutations with presumed negligible impacts on fitness, such as silent nucleotide substitutions. If true, analysis of silent sites would provide a window into the process of mutation. Despite claims to the contrary, I show that neighbouring genes have similar rates of silent evolution. If silent sites evolve neutrally, this pattern reflects regional variation in the mutation rate across the genome. But are silent sites neutrally evolving? I provide multiple lines of evidence that suggest not. In rodents, although two classes of silent sites, introns and synonymous sites, evolve at similar rates, the process of evolution is different between the two. After eliminating conserved intronic sequence, I observe a preference for cytosine at synonymous third sites of codons. I then investigate two models by which synonymous sites might be functional. First, in the mouse lineage, the cytosine preference at third sites and the avoidance of mutations at some synonymous sites can potentially be explained by selection on mRNA secondary structure to promote optimal stability. Second, in humans, biases for particular codons over their synonyms increases near intron-exon junctions, which largely reflects the presence of exonic splicing enhancers. I show that the effect of purifying selection on putative enhancers in mammals leads to a reduction in estimates of the mutation rate.

Chapter 1. **Introduction**

Chance, choice and change in molecular evolution

"multiply, vary, let the strongest live and the weakest die."
(Darwin 1859)

"Darwinism is so well established that it is difficult to think of evolution except in terms of selection for desirable characteristics and advantageous alleles."
(King & Jukes 1969)

While the average person should be able to explain how 'survival of the fittest' can drive evolution, most would be hard-pressed to describe how evolutionary change could occur in the absence of, for example, choice between potential mates. Indeed, before the advent of molecular data, most evolutionary biologists gave little thought to whether natural selection was strong enough act upon all changes at the DNA level. It is for this very reason that King and Jukes infamously entitled their 1969 paper "Non-Darwinian evolution". The issue remains just as relevant today, as the 'intelligent design' branch of creationism attempts to document 'holes' in evolutionary theory by ignoring the importance of chance in molecular evolution (e.g. Behe & Snoke 2004; c.f. Lynch 2005).

At the time that *The Origin of Species* was published, the mechanism by which organisms "vary" was not yet known. Although Mendelian segregation and inheritance later revealed that swapping of segments between parental chromosomes could yield different combinations of alleles (in sexual species), it did not explain how an organism (particularly those that reproduce asexually) could possess a trait that was not observed in any of its ancestors. Of course we now know that mutation of the DNA in germline cells is the major force that adds novel variants to the gene/sequence pool.

When a new mutation arises within a population, whether it spreads to the extent that it is possessed by all individuals (fixation) through choice or chance is dependent on whether the mutation alters the host organism's fitness and whether selection is effective enough to 'notice' the change (Kreitman 1996). The fate of the mutation will only be determined by natural selection if both of these criteria are met, in which case mutations that are detrimental to the fitness of an individual may be purged from the gene pool by purifying (negative) selection, whereas beneficial mutations may spread to fixation through positive (Darwinian) selection. Conversely, if neither of the conditions is met, the probability that the mutation will become fixed may be determined by chance (random genetic drift).

The neutral theory of molecular evolution, the name by which this latter idea has become known, has been one of the dominant models for microevolutionary change ever since its formulation (Kimura 1968; King & Jukes 1969). According to the strictly neutral theory (Kimura 1983), a new mutation can be selectively 'neutral' if it has absolutely no effect on fitness. If such mutations occur at a rate μ per haploid genome per generation, then each generation there must be $2N\mu$ new neutral mutations in a diploid population of size N. Random fluctuations in allele frequency (drift) permit the new mutation to go up or down in frequency. The chance that the mutation will become fixed in a population is $1/2N$.

The fate of a new mutation can also be determined by chance when the mutation has only a very small fitness effect. Under this second scenario, the nearly-neutral theory (Ohta & Gillespie 1996), a mutation will be neutral if its selective disadvantage (s) is small compared to the effective population size, N_e (so that $s<<1/2Ne$, Kreitman 1996). The new mutation is 'effectively neutral' in that the fixation rate is so close to μ as to make no difference. By contrast, if a mutation is 'slightly deleterious', it can be opposed by selection if the fitness effect is larger or the population size smaller (with $s \approx 1/2N_e$), while still allowing fixation to occur at some measurable rate, i.e. a fixation rate less than μ. If the mutation is even more deleterious ($s>>1/2\ N_e$), then the mutation will not reach fixation. Mutations that cause evident disease are just the more extreme examples of those incapable of reaching fixation.

Mammals are of particular interest in this context because, from neutral theory, whether selection is strong enough to exert an influence over the fate of a new mutation is dependent on the effective population size. Organisms such as flies live in populations containing millions of individuals, but $N_e<<10^6$ in mammals (Keightley, Lercher & Eyre-Walker 2005). Therefore, in the latter group, chance is expected to play a major role in determining the fate of mutations of small fitness effect. A mutation in a fly could be slightly deleterious ($s \approx 1/2\ N_e$), while one of the same fitness in a mammal could be effectively neutral ($s<1/2N_e$). The nearly-neutral theory correctly predicts lower levels of selective constraint in small populations (Keightley & Eyre-Walker 2000).

Amino acid sequences, still considered by most biologists to be the smallest units on which selection can act, are not immune to the influence of neutral processes. Based on the observation that the rate at which genes evolve was unexpectedly high, Kimura suggested that most mutations have no impact on an organism's fitness and so spread to fixation through drift. He argued that the rate of protein evolution was such that, if all differences between species were owing to selection, the total amount of death due to natural selection (Haldane 1957) would be improbably high (Kimura 1968). Later, the arrival of protein electrophoresis data implied that polymorphism at

the amino acid level was common (Lewontin 1974), something that was not predicted by selectionist population genetics, but was expected under neutral theory.

While high polymorphism levels and nucleotide substitution rates provide strong support for neutral theory, neutrality by itself cannot explain all of the observed patterns in molecular evolution. For example, species with large populations should exhibit much higher levels of polymorphism than small populations, but this is not observed (Lewontin 1974). Neutral theory also predicts that mutations occur on a regular basis, so that DNA evolution becomes a molecular clock that ticks at a constant rate. However, this too does not appear to be the case (e.g. Cutler 2000).

Even though the above evidence makes it clear that protein evolution is not exclusively neutral, it should be noted that proteins are able to tolerate a wide assortment of amino acid changes (Kimura 1983). About 66% of single amino acid substitutions in human 3-methyladenine DNA glycosylase, for example, do not hinder enzymatic function (Guo, Choe & Loeb 2004). Often the replacement residue possesses similar chemical properties or occurs at a site that does not alter protein structure or functionality (Bowie et al. 1990). Nonetheless, even for the highly conserved catalytic core regions of some proteins, approximately one-third of amino acid sites can tolerate substitutions (e.g. Guo, Choe & Loeb 2004).

To this point, I have followed a protein-centric perspective, primarily because the majority of biologists still hold the view that single nucleotide substitutions (point mutations) are only of significance if they lead to amino acid replacements. Is this a justified assumption? The most monstrous of diseases affecting human morphology are often brought about by mutations that change proteins (Leroi 2003). Based on such observations, it seems intuitively reasonable that sites where mutations do not alter amino acid sequence are relatively unimportant. Indeed, these sites are termed 'silent' based on the assumption that such mutations are phenotypically undetectable and evolutionarily inconsequential.

The silence of the genes

Even before whole genomes had been completely sequenced, it was already known that mammalian DNA consists mainly of silent sites. As only 1.2% of the human genome encodes proteins (Collins et al. 2004), the vast majority are silent mutations. While intergenic DNA can contain functional elements under selection (e.g. Keightley, Lercher & Eyre-Walker 2005), in this thesis I concentrate on evolution at silent sites within genes: intronic sequence and synonymous sites.

With the recent availability of complete genome sequences (Lander et al. 2001; Waterston et al. 2002; Gibbs et al. 2004; Mikkelsen et al. 2005) it is now possible to survey patterns of molecular evolution at the genomic scale (Wolfe & Li 2003). To

study the fate of a new mutation, particularly one that is assumed to have a negligible or zero fitness effect, one must first investigate how mutations arise. For instance, when a point mutation hits a genome, is the location at which it occurs entirely random, or are there certain regions that are more predisposed to being mutated? If the latter is true, why might this be?

In order to address these questions, which I tackle in Part I, one needs to assay the mutation rate across the genome. In mammals, unfortunately, the frequency of spontaneous point mutation is only 10^{-8} per base pair per generation (Drake et al. 1998) and generation times are typically long (e.g. 25 years for hominids, Eyre-Walker & Keightley 1999). Mutation rates are therefore measured indirectly, by counting the number of differences between pairs of orthologous sequences. In such pairwise comparisons, one must assume that a given substitution has already been fixed within the two lineages, hence it is essential to know that the change has been fixed by drift rather than selection. Under the assumption that silent sites are not functional, their rates of evolution can provide a measure of the mutation rate.

Synonymous sites within coding exons occur by virtue of the degeneracy of the genetic code (61 codons specify 20 amino acids), whereas introns, the long stretches of non-coding DNA that lie between the exons, make up about 95% of nucleotides within a gene (Lander et al. 2001). In mammals, both of these classes of silent sites are typically presumed to be non-functional and hence evolve neutrally (e.g. introns, Chang et al. 1994; synonymous sites, Eyre-Walker & Keightley 1999).

Part I investigates whether there is variation in rates of evolution between autosomal regions of mammalian genomes. While several studies have reported that silent site evolution differs between genes (regional variation, e.g. Wolfe, Sharp & Li 1989) and that genes in close proximity evolve at similar rates (local similarity, e.g. Matassi, Sharp & Gautier 1999), some authors maintain that the mutation rate is homogeneous across the genome and that these effects are artefacts caused by the analysis of genes where there is "disparity" in the patterns of evolution between the orthologues in each gene pair (Kumar & Subramanian 2002). In Chapter 2, I ask whether these effects remain following the exclusion of disparately-evolving genes. Our results show that local similarity is a genuine genomic pattern, and that it does not appear to be a consequence of the clustering of genes with similar expression profiles, but may be caused by the mutagenic effects of recombination at meiosis (e.g. Hellmann et al. 2003).

Differences between genes in the rates of silent site evolution may also reflect variable selective constraints (Kimura 1983). As I have already noted above, this explanation has long been considered unlikely because silent mutations do not alter protein sequence. If, however, silent sites are functional, it is possible that silent mutations might affect fitness and so may not evolve neutrally.

Part II examines whether silent sites within murid genes are subject to selection. Neutral theory predicts that the process of evolution should be similar in different classes of silent sites. In Chapter 3, I ask whether introns and exonic four-fold degenerate (synonymous) sites evolve at the same rate. Although I find that the rates of nucleotide substitution are roughly the same, the patterns of substitution are not. Most notably, I observe that cytosine is in excess and unusually stable at four-fold sites, suggesting that there is a bias for usage of certain synonymous codons.

Codon usage bias

Observing that synonymous sites are under selection is an important discovery for mammals, but would not seem so surprising if seen in other taxa. In the early days of the neutral theory, while neutralists argued that the fate of synonymous mutations should be dictated by drift (King & Jukes 1969; Kimura 1977), selectionists countered that, at least in principle, synonymous sites could have functional roles (Clarke 1970). It was not until the early 1980s, however, that evidence emerged for why selection should act at synonymous sites.

In a diverse range of organisms, from bacteria through to plants, yeast, fly and worm, the usage of synonymous codons is biased to maximise the rate of protein synthesis (Ikemura 1985; Akashi & Eyre-Walker 1998; Duret 2002; Wright et al. 2004). This is possible when, for any given set of synonymous codons, the relevant iso-acceptor tRNAs are not equally abundant. If tRNA abundances are skewed and selection favours rapid translation, there might be a pressure to employ the codon that matches the most abundant tRNA. This model predicts that for any given amino acid there is a 'best' (optimal) codon, defined by the skew in tRNA usage, hence also there must exist a set of codons that should be preferred if translation rate is to be maximised. Another prediction is that the bias to favour optimal codons should be most pronounced in highly expressed genes and that experimentally adjusted codon usage should affect expression rates.

In mammals, it is still unclear whether selection for translational efficiency occurs. While some authors find that tRNA abundance (assayed by the copy number of tRNA genes) does not correspond to biased codon usage (Kanaya et al. 2001; Duret 2002; dos Reis, Savva & Wernisch 2004), others report the opposite and propose that a set of optimal codons exists (Comeron 2004). Some data supports a weak relationship between gene expression and codon usage in humans (Urrutia & Hurst 2003; Comeron 2004; Lavner & Kotlar 2005). Similarly, adjusting codon usage can affect net expression levels (Levy et al. 1996; Zolotukhin et al. 1996; Kim, Oh & Lee 1997). Most of these experimental results do not always directly demonstrate that it is translation rate that modulates any effect, however.

Support for the notion that synonymous mutations in mammals are different is also based on the finding that the dominant factor dictating codon usage in mammals is the isochore effect (Eyre-Walker 1991; Sharp et al. 1995; Smith & Hurst 1999). Isochores are large (>300 kb) domains of relatively homogenous guanine+cytosine (GC) content (Bernardi et al. 1985). For a given gene, by far the strongest predictor of nucleotide content at synonymous sites and codon usage bias is the nucleotide content of the isochore (Eyre-Walker & Hurst 2001), i.e. of the flanking non-coding DNA. The underlying cause of isochoric structure remains uncertain (Eyre-Walker & Hurst 2001), but recent evidence (Eyre-Walker 1999; Duret et al. 2002; Lercher et al. 2002) suggests that this too is not simply a neutral process (Galtier et al. 2001; Galtier 2003; Meunier & Duret 2004).

Substrates of selection at synonymous sites

Although the finding that synonymous sites might be under selection goes against conventional wisdom (Sharp et al. 1995), it suggests that isochore effects alone cannot adequately explain synonymous codon usage in mammals. The evidence that I present in Chapter 3 can be considered to be an 'indirect' test for selection in that it can only detect deviations from neutral expectations. It does not, however, provide an explanation for how selection might operate.

Part III collects a series of direct tests for selection. Each test is 'direct' in the sense that a specific mechanistic model by which selection might act is examined. I demonstrate that synonymous codon choice may be important at different stages of gene expression, including post-transcriptional trafficking (Chapter 4) and during pre-mRNA processing (Chapters 5 and 6).

In Chapter 4, I provide evidence that selection on mRNA secondary structure to promote optimum mRNA stability can explain why C is preferred at four-fold sites in rodents (Chapter 3). It also shows that had the synonymous mutations observed in the mouse lineage occurred elsewhere, transcripts would have been less stable.

Chapter 5 investigates two models of selection for efficient splicing, to avoid potential cryptic splice sites or to prefer motifs that enhance splicing. Increased bias in codon usage near intron-exon junctions is expected under both models, so I identify and test the discriminating predictions made by the two models. I find strong support for codons being preferred because they may be exonic splicing enhancers (ESEs).

Following on from this, Chapter 6 asks whether synonymous mutations are selected against in ESEs. I find that these elements are indeed under purifying selection, but that the reduction in synonymous evolution is unlikely to lead to a large underestimate of the genomic mutation rate measured from the synonymous substitution rate.

Chapter 7 is a review of the evidence for selection on synonymous mutations. I apologise that, by necessity, it repeats some of the material in this chapter. As mentioned above, because mammalian population sizes are small, mutations must have a significant impact on fitness in order for selection to overcome drift. The review provides numerous examples of how disrupting the functions of synonymous sites can cause disease. At the end, it describes some of the implications for the finding that synonymous sites do not evolve neutrally. Chapter 8 contains some concluding remarks and future perspectives on this subject.

References

Akashi, H., & Eyre-Walker, A. (1998) Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8**: 688-693.

Behe, M. J., & Snoke, D. W. (2004) Simulating evolution by gene duplication of protein features that require multiple amino acid residues. *Protein Sci.* **13**: 2651-2664.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunierrotival, M. & Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science* **228**: 953-958.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990) Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* **247**: 1306-1310.

Chang, B. H. J., Shimmin, L. C., Shyue, S. K., Hewett-Emmett, D. & Li, W.-H. (1994) Weak male-driven molecular evolution in rodents. *Proc. Natl Acad. Sci. USA* **91**: 827-831.

Clarke, B. (1970) Darwinian evolution of proteins. *Science* **168**: 1009-1011.

Collins, F. S., Lander, E. S., Rogers, J. & Waterston, R. H. (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931-945.

Comeron, J. M. (2004) Selective and mutational patterns associated with gene expression in humans: Influences on synonymous composition and intron presence. *Genetics* **167**: 1293-1304.

Cutler, D. J. (2000) Understanding the overdispersed molecular clock. *Genetics* **154**: 1403-1417.

Darwin, C. (1859) *The Origin of Species*. John Murray, London.

dos Reis, M., Savva, R. & Wernisch, L. (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**: 5036-5044.

Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998) Rates of spontaneous mutation. *Genetics* **148**: 1667-1686.

Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**: 640-649.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D. & Galtier, N. (2002) Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162**: 1837-1847.

Eyre-Walker, A. (1991) An analysis of codon usage in mammals: selection or mutation bias? *J. Mol. Evol.* **33**: 442-449.

Eyre-Walker, A. (1999) Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics* **152**: 675-683.

Eyre-Walker, A., & Keightley, P. D. (1999) High genomic deleterious mutation rates in hominids. *Nature* **397**: 344-347.

Eyre-Walker, A., & Hurst, L. D. (2001) The evolution of isochores. *Nat. Rev. Genet.* **2**: 549-555.

Galtier, N., Piganeau, G., Mouchiroud, D. & Duret, L. (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159**: 907-911.

Galtier, N. (2003) Gene conversion drives GC content evolution in mammalian histones. *Trends Genet.* **19**: 65-68.

Gibbs, R. A. et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**: 493-521.

Guo, H. H., Choe, J. & Loeb, L. A. (2004) Protein tolerance to random amino acid change. *Proc. Natl Acad. Sci. USA* **101**: 9205-9210.

Haldane, J. B. S. (1957) The cost of natural selection. *J. Genet.* **55**: 511-524.

Hellmann, I., Ebersberger, I., Ptak, S. E., Paabo, S. & Przeworski, M. (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**: 1527-1535.

Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**: 13-34.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. & Ikemura, T. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.* **53**: 290-298.

Keightley, P. D., & Eyre-Walker, A. (2000) Deleterious mutations and the evolution of sex. *Science* **290**: 331-333.

Keightley, P. D., Lercher, M. J. & Eyre-Walker, A. (2005) Evidence for Widespread Degradation of Gene Control Regions in Hominid Genomes. *PLoS Biol.* **3**: e42.

Kim, C. H., Oh, Y. & Lee, T. H. (1997) Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells. *Gene* **199**: 293-301.

Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* **217**: 624-626.

Kimura, M. (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**: 275-276.

Kimura, M. (1983) *The Neutral Theory of Evolution*. Cambridge University Press, Cambridge.

King, J. L., & Jukes, T. H. (1969) Non-Darwinian evolution. *Science* **164**: 788-798.

Kreitman, M. (1996) The neutral theory is dead - long live the neutral theory. *Bioessays* **18**: 678-683.

Kumar, S., & Subramanian, S. (2002) Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99**: 803-808.

Lander, E. S. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Lavner, Y., & Kotlar, D. (2005) Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene* **345**: 127-138.

Lercher, M. J., Smith, N. G. C., Eyre-Walker, A. & Hurst, L. D. (2002) The evolution of isochores: evidence from SNP frequency distributions. *Genetics* **162**: 1805-1810.

Leroi, A. M. (2003) *Mutants*. Viking Penguin, New York.

Levy, J. P., Muldoon, R. R., Zolotukhin, S. & Link, C. J., Jr. (1996) Retroviral transfer and expression of a humanized, red-shifted green fluorescent protein gene into human tumor cells. *Nat. Biotech.* **14**: 610-614.

Lewontin, R. C. (1974) *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Lynch, M. (2005) Simple evolutionary pathways to complex proteins. *Protein Sci.* **14**: 2217-2225; discussion 2226-2217.

Matassi, G., Sharp, P. M. & Gautier, C. (1999) Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**: 786-791.

Meunier, J., & Duret, L. (2004) Recombination Drives the Evolution of GC-Content in the Human Genome. *Mol. Biol. Evol.* **21**: 984-990.

Mikkelsen, T. S. et al. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. **437**: 69-87.

Ohta, T., & Gillespie, J. H. (1996) Development of neutral and nearly neutral theories. *Theor. Pop. Biol.* **49**: 128-142.

Sharp, P. M., Averof, M., Lloyd, A. T., Matassi, G. & Peden, J. F. (1995) DNA-sequence evolution: the sounds of silence. *Phil. Trans. R. Soc. Lond. B* **349**: 241-247.

Smith, N. G. C., & Hurst, L. D. (1999) The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? *Genetics* **152**: 661-673.

Urrutia, A. O., & Hurst, L. D. (2003) The signature of selection mediated by expression on human genes. *Genome Res.* **13**: 2260-2264.

Waterston, R. H. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Wolfe, K. H., Sharp, P. M. & Li, W. H. (1989) Mutation rates differ among regions of the mammalian genome. *Nature* **337**: 283-285.

Wolfe, K. H., & Li, W. H. (2003) Molecular evolution meets the genomics revolution. *Nat. Genet.* **33**: 255-265.

Wright, S. I., Yau, C. B., Looseley, M. & Meyers, B. C. (2004) Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. *Mol. Biol. Evol.* **21**: 1719-1726.

Zolotukhin, S., Potter, M., Hauswirth, W. W., Guy, J. & Muzyczka, N. (1996) A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. *J. Virol.* **70**: 4646-4654.

Part I. **Variation in evolutionary rates across the genome**

Spontaneous point mutations arise from errors in DNA replication (that are left uncorrected) and occasionally on exposure to mutagenic chemicals and radiation (Maki 2002). Were these the only factors that dictate the rate of point mutation, one might expect the mutation rate to be uniform across the genome.

To test whether this is true, one assays the rate of nucleotide substitution at sites that are presumed to evolve neutrally. There are advantages and disadvantages to using any class of site (Ellegren, Smith & Webster 2003). In mammals, it has long been assumed that synonymous sites are neutral (Sharp et al. 1995; Duret 2002) and are thus good candidates for representing the mutation rate (Kumar & Subramanian 2002).

Kumar and Subramanian (2002) argued that reports (Wolfe, Sharp & Li 1989; Matassi, Sharp & Gautier 1999; Lercher, Williams & Hurst 2001) that the underlying mutation rate differs among genomic regions (Filipski 1988; Ellegren, Smith & Webster 2003) were unreliable, for several reasons. Most notably, they suggested that regional variation along autosomes (and its corollary, local similarity) is an artefact of the analysis of gene pairs where there is "disparity" in the substitution patterns experienced by each orthologue. Such patterns can result from a gene moving to a new chromosomal region, whereby there may be "amelioration" of its nucleotide content to that of its new genomic location. Consequently, the authors suggested that all genes experience the same underlying mutation rate, that synonymous sites evolve neutrally and that all between-gene variation in evolutionary rate is attributable to estimation error owing to differences in length of sequence.

In Chapter 2, I first ask whether local similarity in rates of evolution is owing to the inclusion of disparate gene pairs. Using both four-fold sites and introns, I find that this is not the case. Even allowing for disparity, neighbouring genes have more similar synonymous rates of evolution than expected by chance. Introns from the same gene also show correlated evolution, including in the rate of insertions/deletions (indels), which should not be susceptible to disparity. That introns from the same gene evolve at similar rates is consistent with investigations of regional variation derived from other forms of non-coding DNA (Chen & Li 2001; Ebersberger et al. 2002; Smith, Webster & Ellegren 2002; Waterston et al. 2002; Hardison et al. 2003; Malcom, Wyckoff & Lahn 2003). The discrepancy with the results of Kumar and Subramanian (2002) appear to be due to their use of a method that is not powerful enough to detect the weak effects of local similarity.

Following the rejection of the notion that local similarity is an artefact, I then go on to investigate the cause of the effect. It does not appear to be due to the observed clustering of genes expressed in the germline, but at least for the rate of synonymous evolution, may be due to the mutagenic effects of meiotic recombination (Perry & Ashworth 1999; Lercher & Hurst 2002; Filatov & Gerrard 2003; Hellmann et al. 2003).

Note that autosomes and sex chromosomes spend different amounts of time in the male and female germlines (Shimmin, Chang & Li 1993) and hence are affected by different mutational forces (e.g. McVean & Hurst 1997; Ebersberger et al. 2002). Consequently, Chapter 2 is only concerned with analysing autosomes.

References

Chen, F. C., & Li, W. H. (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**: 444-456.

Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**: 640-649.

Ebersberger, I., Metzler, D., Schwarz, C. & Paabo, S. (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70**: 1490-1497.

Ellegren, H., Smith, N. G. & Webster, M. T. (2003) Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* **13**: 562-568.

Filatov, D. A., & Gerrard, D. T. (2003) High Mutation rates in human and ape pseudoautosomal genes. *Gene* **317**: 67-77.

Filipski, J. (1988) Why the rate of silent codon substitution is variable within a vertebrates's genome. *J. Theor. Biol.* **134**: 159-164.

Hardison, R. C. et al. (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13**: 13-26.

Hellmann, I., Ebersberger, I., Ptak, S. E., Paabo, S. & Przeworski, M. (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**: 1527-1535.

Kumar, S., & Subramanian, S. (2002) Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99**: 803-808.

Lercher, M. J., Williams, E. J. B. & Hurst, L. D. (2001) Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* **18**: 2032-2039.

Lercher, M. J., & Hurst, L. D. (2002) Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18**: 337-340.

Maki, H. (2002) Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. *Annual Reviews of Genetics* **36**: 279-303.

Malcom, C. M., Wyckoff, G. J. & Lahn, B. T. (2003) Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol. Biol. Evol.* **20**: 1633-1641.

Matassi, G., Sharp, P. M. & Gautier, C. (1999) Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**: 786-791.

McVean, G. T., & Hurst, L. D. (1997) Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* **386**: 388-392.

Perry, J., & Ashworth, A. (1999) Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* **9**: 987-989.

Sharp, P. M., Averof, M., Lloyd, A. T., Matassi, G. & Peden, J. F. (1995) DNA-sequence evolution: the sounds of silence. *Phil. Trans. R. Soc. Lond. B* **349**: 241-247.

Shimmin, L. C., Chang, B. H. & Li, W.-H. (1993) Male-driven evolution of DNA sequences. *Nature* **362**: 745-747.

Smith, N. G. C., Webster, M. T. & Ellegren, H. (2002) Deterministic mutation rate variation in the human genome. *Genome Res.* **12**: 1350-1356.

Waterston, R. H. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Wolfe, K. H., Sharp, P. M. & Li, W. H. (1989) Mutation rates differ among regions of the mammalian genome. *Nature* **337**: 283-285.

Chapter 2. **Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile**

Martin J. Lercher, Jean-Vincent Chamary & Laurence D. Hurst

Genome Research (2004) **14**: 1002-1013

Genomic Regionality in Rates of Evolution Is Not Explained by Clustering of Genes of Comparable Expression Profile

Martin J. Lercher, Jean-Vincent Chamary, and Laurence D. Hurst[1]

Department of Biology and Biochemistry, University of Bath, Bath, BA2 7AY, United Kingdom

In mammalian genomes, linked genes show similar rates of evolution, both at fourfold degenerate synonymous sites (K_4) and at nonsynonymous sites (K_A). Although it has been suggested that the local similarity in the synonymous substitution rate is an artifact caused by the inclusion of disparately evolving gene pairs, we demonstrate here that this is not the case: after removal of disparately evolving genes, both (1) linked genes and (2) introns from the same gene have more similar silent substitution rates than expected by chance. What causes the local similarity in both synonymous and nonsynonymous substitution rates? One class of hypotheses argues that both may be related to the observed clustering of genes of comparable expression profile. We investigate these hypotheses using substitution rates from both human–mouse and mouse–rat comparisons, and employing three different methods to assay expression parameters. Although we confirm a negative correlation of expression breadth with both K_4 and K_A, we find no evidence that clustering of similarly expressed genes explains the clustering of genes of comparable substitution rates. If gene expression is not responsible, what about other causes? At least in the human–mouse comparison, the local similarity in K_A can be explained by the covariation of K_A and K_4. As regards K_4, our results appear consistent with the notion that local similarity is due to processes associated with meiotic recombination.

[Supplemental material is available online at www.genome.org.]

In mammals, it is claimed that the rates of both protein sequence evolution (Williams and Hurst 2000; Lercher et al. 2001) and synonymous nucleotide change (Casane et al. 1997; Matassi et al. 1999; Nachman and Crowell 2000; Lercher et al. 2001; Smith et al. 2002; Yi et al. 2002; Hardison et al. 2003) show local clustering, with neighboring regions evolving at similar rates. However, other authors claim that these results, at least as regards synonymous nucleotide changes, are nothing more than methodological artifacts (Kumar and Subramanian 2002). Here we ask two questions. First, is the local similarity in synonymous substitution rates real? We show that it is. Given this, we then ask why linked genes might have similar synonymous and nonsynonymous substitution rates. In particular, we examine the hypothesis that transcriptional activity provides a possible mechanistic basis for both clustering phenomena (Hurst and Eyre-Walker 2000; Williams and Hurst 2002; Hardison et al. 2003). A priori, a coupling with transcriptional activity is an attractive hypothesis, as genes with comparable expression profile cluster (Caron et al. 2001; Lercher et al. 2002b; Lercher et al. 2003; Versteeg et al. 2003) and expression parameters are related to substitution rates (Duret and Mouchiroud 2000). We show here that transcriptional activity appears not to be an important variable underpinning local similarity of rates of evolution. Finally, we briefly ask what else might then explain the clustering. As the extent of local similarity appears different in the mouse–rat comparison and in the human–mouse comparison (Lercher et al. 2001), we analyze both.

Is Local Similarity an Artifact?

Kumar and Subramanian (2002) argue that all previous findings of local similarity in synonymous substitution rates are invalid, owing to a methodological problem in how one estimates substitution rates. Classical methods (e.g., Tamura and Nei 1993) for estimating substitution rates from pairwise alignments assume that the process of molecular evolution is the same in two sequences since the time of common ancestry. If this assumption is not upheld, the methods will provide biased estimates. Movement of gene clusters into a new environment (e.g., by chromosome rearrangements) can result in the nucleotide composition of the genes "ameliorating" to their new location (Kumar and Subramanian 2002). The inclusion of such sequences with disparate substitution patterns would bias results, leading to overestimates of the actual point substitution rates in the translocated gene cluster (Kumar and Subramanian 2002). Kumar and Gadagkar (2001) developed a disparity test to diagnose such heterogeneity, and suggested that only those alignments passing the test should be employed to investigate mutation rates. Significantly, 46% of human–mouse alignments fail the disparity test, whereas only 8% of mouse–rat alignments are deviant (Kumar and Subramanian 2002).This alone might then explain the observation (Lercher et al. 2001) that local similarity in the synonymous substitution rate is weaker in the mouse–rat than in the human–rodent comparison.

By examining the relationship between the difference in the synonymous substitution rate between two genes and the physical distance between the genes, Kumar and Subramanian (2002) then argued that, in their purified data set, there was no evidence that rates of evolution varied across the genome. Further, the authors argued that all between-gene variation in evolutionary rate is attributable to estimation error owing to differences in length of sequence, and hence that one cannot reject the notion that there is one mutation rate for all autosomal sequences. These and other conclusions from their study have been challenged (e.g., that the global clock rate of synonymous evolution does not differ between mammalian lineages, Yi et al. 2002).

Below, we first confirm that the signals of local similarity

[1]Corresponding author.
E-MAIL l.d.hurst@bath.ac.uk; FAX 44 (0)1225 826779.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1597404.

14:1002–1013 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

examined here are not due to disparate substitution patterns. Additionally, to be more confident that we are examining regionality in nonselective substitution processes, we ask whether, in the mouse–rat analysis, introns from a given gene have more similar rates of evolution than expected by chance. Although thereafter we report results for K_4 only from the subgroup of genes that passed the disparity test, results including all genes are very similar (data not shown). To further minimize the potential effect of disparate evolutionary patterns, we calculate K_4 employing a recently developed scheme that attempts to control for disparity (Tamura and Kumar 2002).

Are Clusters of Germline-Expressed Genes Responsible for Regionality in the Mutation Rate?

The rate of synonymous nucleotide change, assayed at fourfold degenerate sites (K_4), is often assumed to reflect the local mutation rate (but see Eyre-Walker 1999; Smith and Eyre-Walker 2001; Duret et al. 2002; Lercher and Hurst 2002a; Lercher et al. 2002a). Whether more highly expressed genes should be faster or slower evolving at synonymous sites is, however, hard to predict on a priori grounds, as two putative forces have potentially opposite effects. Although there is, for example, evidence that transcription can be mutagenic (Datta and Jinks-Robertson 1995; Aguilera 2002), there are repair mechanisms that are coupled to transcription (Mellon et al. 1987; Selby and Sancar 1993; van Gool et al. 1997). The latter is believed to explain a recently described transcription-associated strand mutational asymmetry in mammals, which has acted to produce a compositional asymmetry, an excess of G+T over A+C on the coding strand, in most genes (Green et al. 2003).

If the two forces (mutation and repair) do not cancel out, a covariation between genic mutation rates and rates of transcription in the germline is to be expected. Prior evidence suggests the possibility of a weak reduction in the synonymous rate of substitution in putatively germline-expressed genes (Duret and Mouchiroud 2000). Given too that it has been shown that highly expressed genes are clustered in the human genome (Caron et al. 2001; Versteeg et al. 2003), transcriptionally mediated variations in the mutation rate could potentially lead to the observation of local similarity.

Are Clusters of Housekeeping Genes Responsible for Regionality in the Rate of Protein Sequence Evolution?

The clustering of highly expressed genes has been shown to be a secondary effect caused by clustering of housekeeping genes (Lercher et al. 2002b). Broadly expressed genes (i.e., those expressed in many tissues, not necessarily at a high rate) are known to evolve at lower rates (Hastings 1996; Duret and Mouchiroud 2000; Williams and Hurst 2002), possibly owing to stronger purifying selection on proteins that have to function in a wide range of different tissues. Thus, such clustering according to breadth of expression might also explain the local similarity in the rate of protein sequence evolution (assayed as K_A, the rate of nonsynonymous substitutions; Williams and Hurst 2002).

Alternative Hypotheses

As regards the above issues, we show that local similarity in both synonymous and nonsynonymous substitution rates is real, but is not explained by the clustering of transcriptionally comparable genes. What other explanations might there be? With regard to the nonsynonymous substitution rate, we ask whether the local similarity is driven by a corresponding local similarity in the synonymous substitution rate. Early claims from the mouse–rat analysis suggested this was not the case (Williams and Hurst 2000), but more recent reanalysis argues to the contrary (Malcom et al. 2003). We return to this issue and ask, why, if local similarity in the nonsynonymous substitution rate is largely owing to underlying variation in the mutation rate, is the effect more pronounced in the vicinity of tissue-specific genes (Williams and Hurst 2002).

Aside from transcription, other possibilities to explain the local variation in the synonymous substitution rate include (1) heterogeneity in the activity of repair enzymes (Matassi et al. 1999), (2) recombination-associated mutational and/or repair hotspots (Perry and Ashworth 1999; Lercher and Hurst 2002b; Filatov and Gerrard 2003; Hellmann et al. 2003), and (3) GC-associated mutation or fixation biases (Lercher et al. 2001; Castresana 2002b; Smith et al. 2002; Yi et al. 2002; Hardison et al. 2003). We examine the last two of these together, as it has been argued that they are not independent (Meunier and Duret 2004).

Methological Issues

Unfortunately, investigations of this nature can suffer a number of problems. First, there is no unambiguously best way to estimate expression parameters (Huminiecki et al. 2003). To have more confidence in any claim that we might wish to make, we use all possible sources of high-throughput data (EST, microarray, and SAGE) so as to test for the robustness of all results. As EST data provide poor representation of expression rates, the latter are estimated from SAGE and microarray data alone. Further, recent evidence suggests that in GC-rich regions, the substitution process may well be affected by both mutation and biased gene conversion (Eyre-Walker 1999; Smith and Eyre-Walker 2001; Duret et al. 2002; Lercher and Hurst 2002a; Lercher et al. 2002a). To distinguish mutation from fixation biases, we also analyze separately GC-poor sequences. Finally, as gene duplications often occur in tandem and as it is possible that the two resulting proteins are under similar purifying selection, neighboring duplicates could also contribute to local similarity in the rate of protein sequence evolution. Although some prior analyses control for the presence of tandem gene duplications (Williams and Hurst 2000, 2002; Lercher et al. 2001), the extent of their potential contribution to local similarity has yet to be evaluated. We report that the effect is substantial even for very distantly related genes, a result which underlines the necessity to elimnate them prior to analysis.

RESULTS AND DISCUSSION

Disparity Is Not Responsible for Local Similarity

Before analyzing any putative causes of local similarity, it is first necessary to establish that such local similarity exists. To test for local similarity, we used a modified version of the method of Lercher et al. (2001). For each gene, we calculated a 'focal average' of the substitution rate, that is, an average over the rates of all other genes within 1 Mb of the focal gene. We denote the correlation coefficient for all data pairs consisting of (1) the rate of the focal gene and (2) the corresponding focal average as the 'focal average correlation', ρ. Thus, the square of ρ estimates what fraction of the variation in the rate K can be explained by comparison to an independent regional average. We estimated statistical significance by a randomization procedure (see Methods). This method was applied to rate estimates of nonsynonymous (K_A) and fourfold degenerate synonymous (K_4) substitutions from orthologous human–mouse and mouse–rat coding sequences, as well as to estimates of intronic point substitution (K_i) and indel (K_{indel}) rates from orthologous mouse–rat introns. Throughout the manuscript, we restrict our analysis of synony-

mous sites to fourfold degenerate third sites; this makes the counting of such sites unambiguous.

Local similarity measured by ρ was significant for synonymous (K_4) and nonsynonymous (K_A) rates, as well as for the intronic substitution (K_I) and indel (K_{indel}) rates (Table 1). Kumar and Subramanian (2002) argued that local similarities in point substitutions may be a methodological artifact, and suggested that genes failing a disparity test should be excluded (Kumar and Gadagkar 2001). We therefore repeated the analysis, this time excluding genes that fail the disparity test ($P_{disparity} < 0.05$, estimated for fourfold degenerate sites and intronic sites, respectively). In approximate agreement with Kumar and Subramanian (2002), we found that 61% of orthologs failed the disparity test in the human–mouse comparison, whereas the corresponding figure for mouse–rat was only 13%. Excluding these disparate genes, we again found significant local similarity for all substitution measures (Table 1, Fig. 1). From the modest reduction in ρ values, we conclude that only a relatively small part of local similarity is caused by the inclusion of disparate genes. To be conservative, all analyzes of K_4 reported below include only nondisparately evolving genes. Very similar results are obtained when including all genes (data not shown). The notion that disparity is not a major cause of local similarity is supported by our finding of a similarly strong local similarity of indel rates, because indels are not expected to suffer from the same estimation problems as point substitutions. Our finding of comparable local similarities for nucleotide substitution and indel rates is consistent with the observation that indels and point substitutions cluster in the same regions (Hardison et al. 2003). Although this suggests that the processes of substitution and insertion/deletion may be mechanistically coupled (Ogata et al. 1996; Hardison et al. 2003), we observe no correlation between the two rates in introns on a within-gene scale (data not shown, c.f. Ogata et al. 1996).

In agreement with the above results, we also found a significant correlation between intronic substitution rates (K_I) and the synonymous substitution rate (K_4) in flanking exons ($r_{Spearman} = 0.17$, $P = 0.030$, both for all genes and for nondisparate genes). The same is reported in the human–mouse comparison after exclusion of fast-evolving sequence (Castresana 2002a). A prior study failed to detect such an effect in the mouse–rat comparison (Hughes and Yeager 1997). It has been suggested that this may be due to a limited sample size (Castresana 2002a), which we supported by a simulated sample size reduction of our data (not shown).

Why does our result differ from that of Kumar and Subramanian (2002), who concluded that controlling for disparity does destroy the signal of local similarity? We believe that the crucial difference lies in the measure of local similarity. Kumar and Subramanian examined the correlation between chromosomal distance and the difference in K_4 across individual, directly neighboring gene pairs. This method has at least two weaknesses. First, it supposes that all of the variation between genes occurs within a chromosomal region and not between chromosomes. If a large part of between-gene variation is actually between chromosomes (Lercher et al. 2001; Castresana 2002b; Ebersberger et al. 2002; Malcom et al. 2003), this method might fail to find any signal. However, the exclusion of between-chromosome effects (by permuting our intronic rates only within the same chromosome) hardly decreased the significance of the local similarity in K_I (data not shown). This confirms the previous notion that a substantial part of local similarity in point substitution rates is independent of chromosomal effects (Lercher et al. 2001).

Further, the method of Kumar and Subramanian may not be sensitive enough for a weak similarity signal. Notably, they considered only individual neighboring gene pairs. As variance in K_4 estimates is dominated by size-dependent noise (Kumar and Subramanian 2002), the value of the difference between two genes will have a large error component; this component may be substantially reduced by the calculation of focal averages. Further, restriction on next neighbors results in very low sample sizes, especially at larger distances. We tested this through an implementation of the method of Kumar and Subramanian (2002). From the human–mouse data set, we first excluded disparately evolving genes. We then calculated the absolute difference in K_4 ($|\Delta K_4|$) between neighboring genes that reside in the same syntenic region. Using a sample size similar to that of Kumar and Subramanian (2002), we confirmed that local similarity is not detectable when comparing mean K_4 across windows of different gene distances (Fig. 1; window size 200 or 500 kb). However, for distances exceeding 1.5 Mb, windows contained inadequate sample sizes well below 50 genes. We suggest therefore that the protocol employed by Kumar and Subramanian (2002) may be too prone to size-dependent error variance to detect what

Table 1. Genomic Estimates of Local Similarity in Evolutionary Rates, Assayed by the Focal Average Correlation ρ

	Human–mouse			Mouse–rat		
	ρ[a]	p[b]	N	ρ[a]	p[b]	N
All genes						
K_A	0.126	<0.0001	4596	0.055	0.0018	4116
K_4	0.334	<0.0001	4284	0.183	<0.0001	3909
K_4^{exCpG}	0.251	<0.0001	4178	0.179	<0.0001	3815
K_I	—	—	—	0.170	<0.0001	541
K_{indel}	—	—	—	0.191	<0.0001	541
Nondisparate genes						
K_A	0.103	0.0003	1568	0.065	0.0010	3554
K_4	0.272	<0.0001	1377	0.166	<0.0001	3380
K_4^{exCpG}	0.189	<0.0001	1292	0.173	0.0002	3298
K_I	—	—	—	0.151	0.0016	469
K_{indel}	—	—	—	0.160	0.0004	469

[a]Focal average correlation = correlation coefficient for data pairs, each consisting of a gene's rate K and the mean of its neighbors within 1 Mb.
[b]Number of equal or higher ρ values found in 10,000 randomly rearranged genomes.

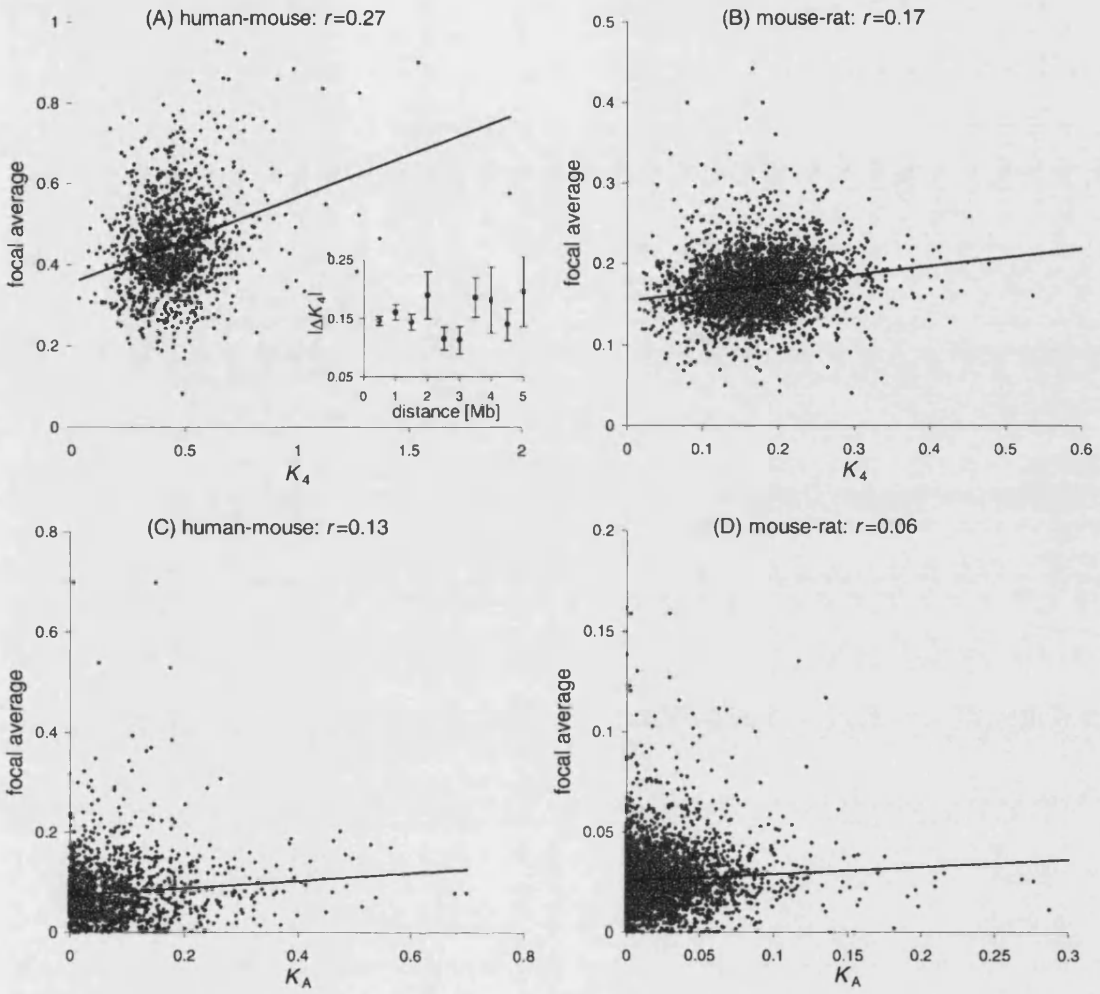


Figure 1 Correlation ρ between substitution rates and focal averages. For K_4 (*A,B*), only nondisparate gene pairs are included. All correlations are highly significant ($P \leq 0.001$). The *inset* in (*A*) shows an alternative measure of local similarity as used by Kumar and Subramanian (2002).

are relatively weak effects, and also failed to take account of between-chromosome variation.

Transcription Affects K_4 but Does Not Explain Local Similarity

Any putative mutational processes associated with transcription are only relevant if occurring in the germline, as only these mutations will reach the next generation. Thus, if transcription does affect the mutation rate, then genes transcribed in the germline should show different substitution rates compared to genes not transcribed in the germline. Unfortunately, we do not have large-scale expression data for mammalian germlines. Although genomic expression data does exist for ovaries and testes (Su et al. 2002), these tissues contain a large number of somatic cells. However, we can gain insight into this issue from analysis of putative housekeeping genes (for definitions see Methods), which should also be expressed in germline cells. In contrast, putative tissue-specific genes (expressed in 0 or 1 of our tissues) are unlikely to be expressed in the germline. As predicted from the hypothesis of transcriptionally induced repair, we found that putative housekeeping genes have significantly lower substitution rates at fourfold degenerate sites (Table 2). However, differences are small, suggesting that any transcriptional effect explains only a small part of the observed variation in substitution rates. Consistent with this difference between tissue-specific and housekeeping genes, and in agreement with previous findings (Duret and Mouchiroud 2000), we also found a significant negative correlation between expression breadth and K_4 (Table 3).

Although the latter effect is weak, we can still ask whether the local similarity in K_4 is owing to the clustering of genes of comparable expression breadth (either because more broadly expressed genes are more likely to be expressed also in germline, or for other unknown reasons). To test this hypothesis, we calculated ρ values for K_4 across all genes with valid expression esti-

Table 2. Difference of Synonymous Substitution Rate (K_4) Between Housekeeping and Tissue-Specific Genes

	Human–mouse					Mouse–rat				
Breadth measure	House-keeping[a]	Tissue-specific[a]	ΔK_4[b]	P[c]	N	House-keeping[a]	Tissue-specific[a]	ΔK_4[b]	P[c]	N
K_4										
EST	0.387	0.502	0.115	0.00077	121	0.161	0.18	0.019	0.00067	463
SAGE	0.411	0.475	0.064	0.0026	394	0.157	0.175	0.018	<0.00001	1928
Microarray	0.387	0.501	0.114	0.00005	103	0.165	0.172	0.007	0.057	789
K_4 residuals from regression on K_A										
EST	0.024	0.055	0.047	121	−0.031	0.0099	0.0012	0.41	463	0.0087
SAGE	−0.002	0	0.51	394	−0.002	0.0092	0.0065	0.032	1928	0.0027
Microarray	0.026	0.048	0.05	103	−0.022	0.0067	−0.0012	0.61	789	0.0079

[a]Average K_4 or average residuals of K_4. Residuals were calculated from expected K_4 values, which were predicted from linear regression of $\log(K_4)$ on $\log(K_A)$ including all genes. Genes were classified as housekeeping/tissue-specific if supported by experiments in both human and mouse for the human–mouse comparison, and by experiments in mouse for the mouse–rat comparison. Only non-disparate gene pairs were included.
[b]Difference in K_4 (or residuals) between tissue-specific and housekeeping averages.
[c]Probability of finding an equal or greater difference in 100,000 randomized genomes.

mates. Statistical significance was estimated by comparison of ρ to 10,000 data sets obtained by randomly permuting the positions of all genes (P_{all}, Table 4). We then repeated the randomization procedure, this time permuting only gene positions within classes of similar expression breadth (P_{group}, Table 4). If local similarity was largely independent of expression breadth, we expect $P_{group} \approx P_{all}$; conversely, if expression breadth determined a large part of the regional variation in K_4, we expect $P_{group} \gg P_{all}$, as randomization in breadth groups would be ineffective in destroying local similarity. From Table 4, we conclude that transcriptional breadth profile per se does not importantly underpin local similarity in K_4.

Might the lower K_4 of housekeeping genes likely to be due to transcription-coupled repair reducing the effective mutation rate of germline-expressed genes? In addition to the failure of a breadth-mediated model to explain local similarity, there are at least two other reasons why this conclusion cannot be accepted at face value. First, it has been suggested that this correlation is a secondary effect caused by the K_A-K_4 correlation, as broadly expressed proteins are known to evolve at slower rates (Duret and Mouchiroud 2000; Williams and Hurst 2002). If in some part K_A drives K_4, by whatever mechanism, then one would need to correct for this. Indeed, when we examined residuals of K_4 from a regression on K_A, the breadth-K_4 correlation disappeared (for each data set; data not shown). Correspondingly, the difference between housekeeping (putatively germline-expressed) and tissue-specific genes was much reduced for the residuals (Table 2).

Given that we are unsure whether K_A does drive K_4 (see below), a better test then is to analyze germline-expressed genes alone, and ask if their rate of expression is a good predictor of K_4. To approximate germline transcription rates for each gene, we calculated the median expression rate for all putative housekeeping genes across all (nongermline) tissues with reported expression. This measure does indeed provide a reasonable approximation of the transcription rate in tissues that are not covered by the experiments. To show this, we performed a benchmarking test, by excluding each of the individual tissues in turn, and calculating the median transcription rate only from the other tissues. This measure is highly correlated with the observed transcription rate for the excluded tissue: For each expression measure, Pearson's r is above 0.5 for the vast majority of all tissues (Supplemental Table S1). However, when correlating this housekeeping gene expression rate with K_4, the results are ambiguous. Although we do find a negative correlation for all measures and data sets (except for human SAGE data), this correlation is nonsignificant in most cases (Table 5; to escape a massive reduction in sample size, results are given for human and mouse expression data separately). The correlations become stronger when restricting the analysis to low-GC genes (Table 5); for the same genes, we also find a stronger breadth-K_4 correlation (Table 3). This is con-

Table 3. Correlation Between Synonymous Substitution Rate (K_4) and Expression Breadth

	Human–mouse			Mouse–rat		
Breadth measure[a]	r[b]	P	N	r[b]	P	N
All genes						
EST	−0.203	<0.00001	1247	−0.063	0.00088	2759
SAGE	−0.125	0.00007	1057	−0.088	<0.00001	2579
Microarray	−0.187	0.00003	596	−0.057	0.0237	1592
GC≤0.5						
EST	−0.341	<0.00001	322	−0.165	0.00011	570
SAGE	−0.242	0.00011	260	−0.202	<0.00001	565
Microarray	−0.339	0.00002	148	−0.090	0.098	340

[a]Breadth of expression was averaged over experiments in human and mouse for the human–mouse comparison, and was obtained from mouse only in the mouse–rat comparison.
[b]Pearson's correlation coefficient between expression breadth and K_4. Only nondisparate gene pairs were included.

Table 4. Effect of Expression Breadth on Significance of Local Similarity (ρ) in K_4

	Human–mouse				Mouse–rat			
Breadth measure[a]	ρ	P_{all}[b]	P_{group}[c]	N	ρ	P_{all}[b]	P_{group}[c]	N
EST	0.290	<0.0001	<0.0001	933	0.188	<0.0001	<0.0001	2544
SAGE	0.323	<0.0001	<0.0001	732	0.177	<0.0001	<0.0001	2351
Microarray	0.506	<0.0001	<0.0001	320	0.173	<0.0001	<0.0001	1317

[a]Breadth of expression was averaged over experiments in human and mouse for the human–mouse comparison, and was obtained from mouse only in the mouse–rat comparison. Only nondisparate gene pairs were included.
[b]P_{all} is the fraction of equal or greater ρ in datasets obtained by randomly permuting all genes.
[c]P_{group} is the fraction of equal or greater ρ in datasets obtained by randomly permuting genes within classes of similar K_4.

sistent with recent reports of a fixation bias in synonymous substitutions in GC-rich genes (Eyre-Walker 1999; Smith and Eyre-Walker 2001; Duret et al. 2002; Lercher and Hurst 2002a; Lercher et al. 2002a); accordingly, GC-depleted genes may be expected to most accurately reflect mutational biases. Conversely, this result suggests that for the most gene-dense regions, characterized by $GC_4 > 0.5$, systematically varying fixation biases may dominate the transcription-coupled mutational biases examined here.

In sum, we have suggestive evidence that synonymous substitution rates of housekeeping genes may be directly affected by rates of expression, consistent with the action of transcription-coupled repair processes in the mammalian germline (Svejstrup 2002) or with stronger purifying selection acting on the most abundantly expressed genes (Urrutia and Hurst 2003). However, the effect on K_4 is at most weak, and is unlikely to contribute much to the observed patterns of local similarity in K_4.

Alternative Explanations for Local Similarity in K_4

As the transcription-mediated model failed to account for local similarity in K_4, we need to examine alternative hypotheses. The rate of synonymous evolution covaries with regional GC content (Smith and Hurst 1999; Bielawski et al. 2000; Hurst and Williams 2000; Castresana 2002b; Smith et al. 2002; Hardison et al. 2003), although the exact form and strength of this relationship is still a matter of debate, and seems to depend on methodology (Hurst and Williams 2000; Bierne and Eyre-Walker 2003). There is also evidence of a positive correlation between K_4 and meiotic recombination rate (Perry and Ashworth 1999; Lercher and Hurst 2002b; Filatov 2003; Filatov and Gerrard 2003; Hellmann et al. 2003), possibly indicating a mutagenic effect of recombination. GC and recombination rates are known to covary (Eyre-Walker 1993; Fullerton et al. 2001), and it has recently been argued that substitutional GC biases are directly associated with recombination events (Meunier and Duret 2004). Both GC and recombination rate are known to fluctuate systematically over megabase-sized regions, suggesting that regionality in K_4 may be a secondary effect of these variations.

We performed a multiple linear regression of K_4 on recombination rate and on three different GC measures for the human–mouse alignments: GC_4 (GC at fourfold degenerate sites, averaged over the aligned sequences); GC_i (intron GC, averaged over the aligned sequences); and $|\Delta GC_i|$ (absolute difference in intron GC between the aligned sequences). We found that $r^2 = 0.104$ of the variation in K_4 can be explained by these four variables, with all variables contributing significantly (t-test, $P < 0.001$ for each variable). When restricting this analysis to nondisparate genes, only GC_4 and recombination rate contribute significantly ($r^2 = 0.077$). We can ask to what extent this covariation contributes to local similarity in K_4, by examining the residuals of the multiple regression. We found that ρ for K_4 is reduced from 0.34 to 0.28 when analyzing the residuals (nondisparate genes: 0.30 to 0.23). Thus, part of the regionality in K_4 can be attributed to effects of GC and recombination rate variation in the human–mouse comparison, even after exclusion of genes exhibiting heterogeneous substitution patterns.

As we were unable to obtain fine-scale recombination rate estimates for rodents, we only tested the influence of the different GC measures on the variation of K_4 in the mouse–rat comparison. In contrast to the above result, the fraction of K_4 variation explained by GC is very low ($r^2 = 0.002$); only GC_4 contributes significantly ($P = 0.016$). This is practically unchanged when restricting the analysis to nondisparate genes ($r^2 = 0.003$). Corre-

Table 5. Correlation Between Expression Rate and K_4 for Putative Housekeeping Genes

	Human–mouse			Mouse–rat		
Rate measure[a]	r	P	N	r	P	N
All genes						
SAGE (human)	0.026	0.761	141	—	—	—
Microarray (human)	−0.134	0.079	172	—	—	—
SAGE (mouse)	−0.068	0.388	166	−0.003	0.96	305
Microarray (mouse)	−0.029	0.684	196	−0.050	0.33	391
$GC_4 \leq 0.5$						
SAGE (human)	−0.446	0.006	37	—	—	—
Microarray (human)	−0.038	0.838	32	—	—	—
SAGE (mouse)	−0.401	0.005	47	−0.089	0.43	80
Microarray (mouse)	−0.137	0.319	55	−0.006	0.96	93

[a]Rate measures were not averaged over human and mouse in this table to retain acceptable sample sizes. Only nondisparate gene pairs were included.

spondingly, local similarity was unchanged when analyzing the multiple regression residuals.

One possible reason for a relationship between K_4 and GC content is the hypermutability of CpG dinucleotides. Indeed, it has been suggested that mammalian isochores (regions of varying GC content) are the historical consequence of varying substitution rates at CpG sites (Arndt et al. 2003). To exclude the influence of these sites, we also calculated K_4^{excCpG} after removing all sites contributing to a CpG dinucleotide in any of the aligned sequences. Local similarity is largely unaffected by this (Table 1), suggesting that a substantial fraction of regional variation in the mutation rate is caused by processes at nonCpG sites. Consistent with this supposition, all results reported in Tables 2–7 are essentially unchanged when replacing K_4 with K_4^{excCpG} (Supplemental Tables S2–S6).

It is interesting to note that we found local similarity in K_4 to be strongest for the subset of genes with high GC (human–mouse, $GC_4 > 0.7$: $\rho = 0.473$, $P < 0.0001$, n = 234, nondisparate only). Regions ('isochores') of high GC are also known to exhibit much more small-scale variation in GC than regions of lower GC content (Nekrutenko and Li 2000). It has been suggested that this variation is caused by biased gene conversion at local hotspots of recombination (Meunier and Duret 2004). Thus, it is conceivable that local similarity in K_4 is indeed caused by recombination-induced effects. Our inability to confirm such a relationship may then simply reflect the fact that recombination rate estimates are regional averages (Kong et al. 2002) and are only a poor predictor of the ancestral recombination events that shaped substitution patterns. This view is consistent with a recent study that found a very strong correlation between human recombination rates and the GC bias of recent nucleotide substitutions (Meunier and Duret 2004).

Breadth of Expression Has, at Most, a Weak Effect on Local K_A Similarity

Before we can assess the contribution of expression-mediated effects on K_A, we must first exclude duplicated genes from the analysis (Williams and Hurst 2000; Lercher et al. 2001). As duplicated genes may often be subject to similar selective constraints, their rates of amino acid substitution will be correlated. Duplicated genes often reside close to each other and may thus contribute to a signal of local similarity. We compared the K_A between the two copies of 472 human–mouse duplicate gene pairs, identified by significant sequence similarity (pairwise blast expectation value $E \leq 0.001$), with at most 1 Mb distance between them on a human autosome. As expected, the K_A values of duplicated genes are strongly correlated ($r = 0.54$). Surprisingly, the same protocol showed a much weaker similarity of duplicate genes in the mouse–rat comparison ($r = 0.26$). It is interesting to note that K_A values are still correlated between very distantly related sequences: When sorting neighboring gene pairs (those within 1 Mb of each other) into subsets of 100 pairs according to pairwise BLAST score, we found that the correlation between K_A values was significantly increased for all sets with expectation values $E < 0.01$ (Lercher et al. 2002b). To test the effect of these correlations on the local K_A similarity, we recalculated ρ, this time excluding all putative duplicates (Table 6). ρ was reduced by ~50% in both the human–mouse and mouse–rat comparisons. Local similarity was still significant in the human–mouse comparison. However, in the mouse–rat comparison, local similarity in K_A became nonsignificant after strict removal of duplicate genes. In the mouse–rat comparison, local K_A similarity (before duplicate exclusion) was enhanced when restricting the analysis to low-GC genes, and was strengthened by further excluding genes with disparate substitution patterns ($\rho = 0.21$). For this subclass of genes, local similarity remained significant even after exclusion of duplicates ($\rho = 0.17$, $P = 0.0023$).

Table 6. Local Similarity in Nonsynonymous Substitution Rate K_A, Including All Genes or Excluding Duplicate Genes

	Human–mouse			Mouse–rat		
	ρ	*P*	*N*	ρ	*P*	*N*
All genes	0.126	<0.0001	4596	0.055	0.0018	4116
Nonduplicates[a]	0.065	0.0002	4515	0.022	0.087	3995

[a]Genes were excluded from the focal average if they exhibited significant sequence similarity to the focal gene (BLAST expectation value *E* <0.02; Lercher et al. 2001).

We then proceeded to analyze the influence of gene expression on K_A. In agreement with previous studies (Duret and Mouchiroud 2000; Williams and Hurst 2002), our analysis confirmed a significant negative correlation between different measures of expression breadth and the rate of protein evolution, K_A (Table 8).Depending on the data sets used, between 2% and 20% of the variation in K_A can be predicted by a gene's expression breadth. Although these numbers appear relatively low compared to Figure 1 in Duret and Mouchiroud (2000), it must be kept in mind that the latter study grouped genes according to similar expression breadth, and thereby filtered out a large proportion of additional variation. When comparing different expression assays, we found that the strongest correlations are consistently seen with EST data; this may be a simple consequence of the large number of tissues for which EST data are available. Further, it is striking that the K_A-breadth correlation is enhanced when restricting the analysis to low-GC genes (Table 8). This is reminiscent of a similar effect for K_4 (Table 3).

Next, we analyzed local similarity (ρ, excluding duplicate genes) for all genes with valid measures of expression and K_A, assessing statistical significance by comparison to 10,000 randomized data sets. To test whether regional variation in expression breadth is responsible for part of the local similarity in K_A, we repeated the randomization procedure, this time permuting gene positions only between genes with similar breadth of expression (Table 9). The local similarity measure ρ in Table 9 is generally low, probably due to the sample-size reductions associated with the expression data sets. Controlling for expression breadth has practically no influence on the estimated statistical significance of ρ ($P_{all} \approx P_{group}$). We conclude that, at most, only a small part of local similarity in K_A can be explained by the covariation with expression breadth, and alternative explanations must be sought.

Alternative Explanations for Local Similarity in K_A

As commonly described (Li 1997; Smith and Hurst 1999), a sizeable part of the variation in the protein sequence rate of evolution (K_A) is predicted by K_4, the substitution rate at fourfold degenerate sites (human–mouse: $r = 0.39$, $P < 0.00001$, n = 4726; mouse–rat: $r = 0.23$, $P < 0.00001$, n = 4092; but see also Bielawski et al. 2000). This covariation has been attributed, in part, to correlated (but not necessarily simultaneous) substitutions between neighboring synonymous and nonsynonymous sites (tandem substitutions; Smith and Hurst 1999; Duret and Mouchiroud 2000). When, for example, mouse–rat genes with no tandem substitutions are analyzed, there is no K_A-K_4 correlation (Smith and Hurst 1999). When following the method of Duret and Mouchiroud (2000), by excluding all fourfold degenerate changes neighboring a substitution at the first site of the next codon, we find

Table 7. Effect of K_4 on Significance of Local Similarity in K_A (Duplicate Genes Excluded)

	Human–mouse				Mouse–rat			
	ρ	P_{all}[a]	P_{group}[b]	N	ρ	P_{all}[a]	P_{group}[b]	N
All genes	0.053	0.0008	0.046	4191	0.026	0.063	0.093	3787
Nondisparate	0.059	0.025	0.21	1312	0.039	0.018	0.029	3264

[a]P_{all} is the number of equal or greater ρ in datasets obtained by randomly permuting all genes.
[b]P_{group} is the number of equal or greater ρ in datasets obtained by randomly permuting genes within classes of similar K_4.

that the correlation is substantially reduced (human–mouse: $r = 0.13$, $P < 0.00001$, n = 4749; mouse–rat: $r = 0.13$, $P < 0.00001$, n = 4085). However, this could be an overestimate of the impact of tandem substitutions: even if neighboring substitutions occur independent of each other, fast-evolving genes will have far more tandem substitutions than slowly evolving genes ($\sim K_A \times K_4$). Thus, we will take out disproportionately more synonymous substitutions for fast-evolving genes than for slowly evolving genes, which of course reduces the correlation between K_A and K_4.

To get an unbiased estimate of the underlying mutation rate excluding any potential tandem substitution effects, we must adjust not just the number of substitutions, but also the number of fourfold degenerate sites that are used to calculate substitutions per site. Thus, we recalculated K_4, this time excluding all fourfold degenerate sites (with or without substitutions) that were followed by a codon with a substitution at its first site. This new K_4' was still correlated with K_A (human–mouse: $r = 0.28$, $P < 0.00001$, n = 4725; mouse–rat: $r = 0.18$, $P < 0.00001$, n = 4064).

In sum, although tandem substitution biases appear to exist [K_4' is on average 3% (human–mouse) or 1.6% (mouse–rat) lower than K_4], these biases cannot fully account for the K_A-K_4 correlation. This finding is consistent with a number of recent analyses, which, in contrast to early reports (Averof et al. 2000), find evidence for only a very low rate of doublet mutations (Silva and Kondrashov 2002; Kondrashov 2003; Smith et al. 2003). In further support of our conclusion that the K_A-K_4 correlation is not due to mechanistic coupling of substitutions of neighboring sites, we also found a significant positive correlation between K_A and K_i (the substitution rate within introns of the same gene, $r_{Spearman} = 0.248$, $P = 0.005$, n = 127).

The strong K_A-K_4 correlation demonstrated above suggests that much of the local similarity in K_A may be explained by the local similarity in K_4. To test this hypothesis, we calculated ρ values for K_A across all genes with valid estimates for K_A and K_4, excluding duplicate genes from the focal averages. Statistical significance was estimated by comparison of ρ to 10,000 data sets obtained by randomly permuting the positions of all genes (P_{all}, Table 7). We then tested for the effect of K_4, by repeating the randomization procedure, this time permuting only gene positions within classes of similar K_4 (P_{group}, Table 7). For the human–mouse comparison, significance of the ρ value was markedly reduced, and became marginally significant (all genes; nonsignificant when ρ was calculated as Spearman's rank correlation coefficient) or nonsignificant (nondisparate genes or GC ≤ 0.5). Thus, after the exclusion of duplicate genes, the majority of local K_A similarity in the human–mouse comparison can be attributed to the genes' synonymous substitution rates.

Consistent with an earlier analysis (Williams and Hurst 2000), we obtained a markedly different result in the mouse–rat comparison (Table 7). Significance of the ρ value hardly depends on randomization protocol ($P_{all} \approx P_{group}$), suggesting that the underlying mutation rate contributes very little to the local similarity among rodents. In fact, local similarity in K_A was nonsignificant for the mouse–rat comparison after removal of duplicate genes (Table 6). What causes this difference between the two species comparisons? If we accept the notion that a coupling of K_A and K_4 is responsible for the local similarity observed in the human–mouse comparison, this suggests that the K_A-K_4 coupling is reduced in rodents. One possible reason may be that the effective population sizes of rodents are larger, and thus fewer amino acid substitutions are effectively neutral; however, further analyses are necessary to resolve this issue.

A Model for the Strength of Local Similarity in K_A

In apparent contrast to our finding that transcription does not have a significant role in local K_A similarity, a previous analysis of mouse–rat orthologs reported that local similarity in K_A is most pronounced in the vicinity of narrowly expressed (tissue-specific) genes (Williams and Hurst 2002). This we have confirmed for our human–mouse data set. When analyzing separately each of five breadth classes, we found significant local similarity only for the most narrowly expressed genes (excluding duplicate genes: ρ = 0.084, $P = 0.010$ including Bonferroni correction for multiple tests, n = 1241). The local similarity estimated for this subset of

Table 8. Correlation Between Expression Breadth and Nonsynonymous Substitution Rate, K_A

	Human–mouse			Mouse–rat		
Breadth measure	r	P	N	r	P	N
All genes						
EST	−0.286	<0.00001	3451	−0.204	<0.00001	3305
SAGE	−0.197	<0.00001	2925	−0.158	<0.00001	3075
Microarray	−0.242	<0.00001	1624	−0.133	<0.00001	1909
Nondisparate & GC≤0.5						
EST	−0.463	<0.00001	322	−0.341	<0.00001	570
SAGE	−0.276	<0.00001	260	−0.200	0.00001	565
Microarray	−0.363	<0.00001	148	−0.225	0.00003	340

Table 9. Effect of Expression Breadth on Significance of Local Similarity in K_A (Duplicate Genes Excluded)

	Human–mouse				Mouse–rat			
Breadth measure[a]	ρ	P_{all}[b]	P_{group}[c]	N	ρ	P_{all}[b]	P_{group}[c]	N
EST	0.047	0.0072	0.016	3115	0.023	0.10	0.069	3029
SAGE	0.041	0.032	0.044	2572	0.041	0.024	0.020	2740
Microarray	−0.037	0.90	0.92	1209	0.038	0.071	0.049	1559

[a]Breadth of expression was averaged over experiments in human and mouse for the human–mouse comparison, and was obtained from mouse only in the mouse–rat comparison.
[b]P_{all} is the number of equal or greater ρ in datasets obtained by randomly permuting all genes.
[c]P_{group} is the number of equal or greater ρ in datasets obtained by randomly permuting genes within classes of similar expression breadth.

putative tissue-specific genes is actually higher than that estimated for the total data set (ρ = 0.065). Does this imply a direct coupling of gene expression with regionality in K_A? We wish to suggest that this might not necessarily be so.

Consider a simple model supposing that the local similarity in K_A (excluding duplicate gene effects) is driven by the local similarity in the mutation rate (Malcom et al. 2003). Let us further assume that K_4 is somehow coupled to the mutation rate, which as noted above need not be true. If one subsamples any given group of genes, the extent to which one will detect local similarity will then depend on the extent to which, within the subsample, K_A is coupled to K_4. Consider now two extremes: (1) a set of proteins that evolve neutrally ($K_A = K_4$), and (2) another set under extreme purifying selection ($K_A = 0$ for all). In the former, K_A and K_4 are perfectly coupled; in the latter there is no coupling. More generally, when we select subsamples under different degrees of purifying selection, then lower selection pressures will correspond to a higher proportion of effectively neutral amino acid substitutions, and thus to stronger K_A-K_4 coupling. If narrowly expressed genes are under weaker purifying selection (as appears to be the case, see above), then we expect a tighter coupling of K_A and K_4, and consequently a stronger signal of local similarity when only these genes are analyzed. Our finding that the K_A-K_4 correlation cannot be accounted for by tandem mutations is of importance for this interpretation, as otherwise one might suppose that K_A drives K_4, not the other way around.

As expected from our model, the K_A-K_4 coupling is strongest in tissue-specific genes: $r = 0.42$, compared to $r = 0.39$ when including all genes (human SAGE data). This observation also strengthens the notion that the coupling is caused by a fraction of effectively neutrally evolving amino acid positions. If alternatively the coupling was due to similar selection pressures on both nonsynonymous and synonymous sites, then the coupling should be stronger when considering the full range of expression profiles.

We applied two additional tests for our model. We first asked whether the observed strength of the K_A-K_4 coupling in randomly drawn subsets of our data predicts the strength of local similarity found within the subset: this is indeed the case ($r_{Spearman} = 0.083$, $P = 0.004$, from examining the dependence of ρ on the correlation coefficient r between K_A and K_4, for 1000 randomly drawn subsets of 1000 genes). Secondly, we compared the quarter of our data set with the highest K_A/K_4 to the quarter exhibiting the lowest K_A/K_4; these groups putatively correspond to genes under low and high selective pressures, respectively. As expected, we found stronger similarity in the genes with higher K_A/K_4 (excluding duplicate genes: ρ = 0.132 vs. 0.084); this group also exhibits stronger K_A-K_4 coupling ($r = 0.62$ vs. 0.51).

Thus, the stronger local similarity for narrowly expressed genes might be explained as a consequence of the stronger K_A-K_4 coupling, which in turn is due to a fraction of sites that evolve effectively neutrally. There is, however, at least one problem with the null model, this being that the local similarity in K_4 extends over many megabases (Smith and Lercher 2002), compared to less than 3 Mb for local similarity in K_A (Lercher et al. 2001). The local similarity in K_4 decreases, however, with increasing distance. It is thus possible that at larger distances secondary effects on K_A are present, but are too small to be detected.

In sum, we have demonstrated that transcription has little to do with establishing local similarity in rates of evolution. Local similarity in K_A appears to be largely due to tandemly duplicated genes, and to the coupling of K_A to the mutation rate for sites encoding amino acids that evolve neutrally. In turn, the local similarity in K_4 may be largely due to recombination-associated effects.

METHODS

Accession numbers of orthologs and associated data used for the analyses are available as online Supplemental data.

Orthologous Coding Sequence Identification

We obtained lists of putative human–mouse and mouse–rat orthologous genes, identified through reciprocal best BLAST hits, from Ensembl (http://www.ensembl.org; Hubbard et al. 2002). If a gene in one species matched more than one gene in the other species, indicating a lineage-specific gene duplication, it was excluded from further analysis. This resulted in primary data sets of 13,015 (human–mouse) and 12,637 (mouse–rat) orthologous gene pairs. We excluded human transcripts without known position on the UCSC November 2002 genome assembly (http://genome.ucsc.edu), and mouse genes without known position on the Ensembl map. As evolutionary forces affecting genes located on the sex chromosomes differ from those affecting autosomal genes, we restrict our analyses to genes located on human or mouse autosomes.

For each gene, we then downloaded all transcripts from Ensembl. We excluded those transcripts where the coding sequence lacked a valid start or stop codon. For each orthologous gene pair, we selected matching transcripts. This was done under the assumption that a large proportion of genes is alternatively spliced, and that two transcripts corresponding to analogous splice forms should have similar lengths. We first searched for the longest pair of transcripts with a length difference of at most 1%; if no transcript pair fulfilled this criterion, we selected the transcript pair most similar in length. If all transcript pairs differed by more than 5% in their length, we discarded the gene. This procedure resulted in sets of 5212 (human–mouse) and 4442 (mouse–rat) transcript pairs, where transcript pairs will generally correspond to analogous splice forms. For the human–mouse comparison, distances were measured between transcription midpoints of genes on the human UCSC November 2002 assem-

bly; for the mouse–rat comparison, we used the Ensembl mouse map (build 30).

Nucleotide Alignments and Evolutionary Distances

Transcript coding sequences were first translated to amino acid sequences. These were aligned using Clustalw (Thompson et al. 1994) with default settings. The amino acid alignments were then used as templates to align the nucleotides. We calculated nonsynonymous distances (K_A) using the method of Li (1993) and the Kimura two-parameter model. The neutral substitution rate was estimated from the distance at fourfold degenerate sites (K_4), using only codons with no changes at other sites. These rates were corrected for multiple hits with a model that accounts for compositional biases and for substitution pattern differences (disparity) between the two sequences (Tamura and Kumar 2002). We defined gene classes of similar K_4 by dividing the ranked data set into 10 equally sized groups.

For intronic substitution rates, only introns located between coding exons were analyzed. Two methods were used for aligning introns: manually (by-eye) and MCALIGN, a stochastic maximum likelihood-based (ML) program that incorporates a Monte Carlo algorithm (http://homepages.ed.ac.uk/eang33/mcinstructions.html). We executed the program using the rodent intron parameters provided. Seven large introns (>7 kb) proved too difficult to align. Our final intron data set consisted of 136 orthologous genes possessing 560 introns. For further details see Chamary and Hurst (2004).

Within introns we excluded the first and last 20 base pairs, as these appear to be subject to purifying selection (Majewski and Ott 2002; Chamary and Hurst 2004). To obtain genic K_i values, intronic substitution rates were weighted according to the number of bases compared per individual intron alignment (Smith and Hurst 1998). The indel rate, K_{indel}, was calculated as the total number of indels per base pair of the alignment. After estimating the K_i per intron by the two alignment methods (manual and from the ML protocol), we defined a conservative set and a liberal set. For any given intron, these two sets contain the alignment that yields the lower or higher K_i respectively, providing two estimates for the rate of evolution of each intron. We here report results only for the conservative set; very similar results were obtained for the liberal set, as well as for an additional set that consists of only slow-evolving regions (data not shown; Castresana 2000; http://www1.imim.es/~castresa/Gblocks/Gblocks.html).

GC Content and Recombination Rates

For each gene, the guanine + cytosine content at fourfold degenerate sites, GC_4, was averaged over the two aligned coding sequences. We further calculated intron GC for each transcript from the compositional difference between mRNA and exon sequences (downloaded from Ensembl), as $GC_{intron} = (GC_{mRNA} \times length_{mRNA} - GC_{exons} \times length_{exons}) / (length_{mRNA} - length_{exons})$. From this, we calculated GC_i^{avg}, the average over the two aligned transcripts, and GC_i^{diff}, the absolute difference between the two transcripts. Recombination rate estimates for humans were obtained from the UCSC genome browser (http://genome.ucsc.edu), and are based on the deCODE data (Kong et al. 2002).

Expression Breadth and Rate

EST Data

Human and mouse Ensembl genes were mapped to NCBI UniGene clusters (Schuler et al. 1996; UniGene build 161, obtained from NCBI at ftp://ncbi.nlm.nih.gov/repository/UniGene) via RefSeq sequence IDs. Only unambiguous pairings were retained. dbEST library accessions for all ESTs mapping to these clusters were extracted from UniGene. For each library mapping to at least 50 UniGene clusters, the associated tissue type was obtained from dbEST annotation (http://ncbi.nlm.nih.gov/dbEST). We kept only libraries based on well defined, nondisease tissue types. Libraries representing the same tissue type were joined, and tissues matching <500 Ensembl genes were excluded. This resulted in a data set containing 14,559 genes expressed in at least one out of 55 tissue types for humans, and in a second set containing 11,418 genes expressed in at least one out of 49 tissue types for mouse. For each gene, breadth of expression was estimated as the fraction of tissues with an observed EST.

SAGE Data

Serial Analysis of Gene Expression (SAGE) data (Velculescu et al. 1995) was obtained from SAGEmap (Lash et al. 2000) at NCBI (ftp://ncbi.nlm.nih.gov/pub/sage). The data sets were curated to avoid possible GC biases in SAGE libraries, following the approach of Margulies et al. (2001), by removing libraries with mean tag GC > 0.5. The resulting SAGE tag/tissue data sets were based on 40 libraries representing 19 nondisease tissues (human), and on 23 libraries representing nine nondisease tissues (mouse). Tag counts for each data set were converted to relative values (cpm, counts per million) after joining all libraries representing the same tissue type. If tags were found only once in one tissue type, we discarded the observation as a likely sequencing error. These data sets were cross-linked to the mRNA sequences in RefSeq (ftp://ncbi.nlm.nih.gov/refseq), by extracting the 3'-most NlaIII and Sau3A SAGE tags for each human and mouse mRNA. These were then cross-linked to Ensembl genes. We disregarded all tags mapping to more than one Ensembl gene, and excluded the associated genes from further analysis. If several tags mapped to the same gene (representing alternative splice forms), we used maximum cpm in each tissue. In human, we obtained reliable SAGE tags for 11,507 genes, with 7285 expressed in at least one tissue. In mouse, we collected expression data for 10,480 genes, of which 5016 were expressed in at least one tissue. For each gene, we calculated breadth of expression as the fraction of tissues with cpm > 0. Rate of expression was defined as cpm in each tissue.

Microarray Data

Normalized microarray expression data based on Affymetrix chips for 7315 human and 5971 mouse genes were obtained from Su et al. (2002). Human data were sorted into 28 nonredundant tissue types, encompassing 63 replicate hybridizations. Mouse data for 45 tissue types were based on 98 replicate hybridizations. For each tissue, the mRNA expression level (termed 'expression rate' to be consistent with SAGE terminology) was estimated as the mean across replicates. Because there is no unambiguous way to distinguish expressed from nonexpressed data in this type of experiment, we based our breadth measure on observed expression rates, as breadth = (average mRNA expression level across tissues) / (maximum mRNA expression level across tissues). Breadth was set to 0 if the mRNA expression level was <50 in all tissues; this low level could be chosen because our method effectively smoothes out experimental error by joining information across tissues. We also tried other breadth measures (such as defining genes with level < 100 as nonexpressed, and genes with level > 200 as expressed; Su et al. 2002), with very similar results (data not shown).

Definition of Breadth Classes

For the analysis in similar breadth classes, we subdivided the data set into classes of width 0.05. As the highest breadth classes contain the lowest numbers of genes, we further joined these until the highest class contained at least 20 genes. For further analysis of subgroups of genes (nondisparate, low-GC), we joined neighboring classes until each class contained at least 10 genes.

All expression assays are biased against genes expressed at low levels, which may not be detected unless very high numbers of ESTs, SAGE tags, or replicate hybridizations are analyzed. For this reason, we expect many cases where absence of gene expression is wrongly inferred from the data. Accordingly, we must allow for false negatives in some tissues when selecting putative housekeeping genes. For microarray data, we treated all mRNA expression levels <100 as nonexpressed, all >200 as expressed, and all other as unknown (Su et al. 2002). We conservatively labeled those genes without nonexpressed tissues as putative

housekeeping genes (human: 8.5% of genes; mouse: 11.1%). Based on this analysis, we estimated that at least 10% of all genes should perform housekeeping functions. For EST and SAGE data, we selected corresponding thresholds of observed expression breadth for putative housekeeping genes: human SAGE, 0.7 (7.6% of genes); mouse SAGE, 0.5 (9.0%); human EST, 0.52 (9.7%); mouse EST, 0.6 (7.9%). In all assays, putative tissue-specific genes were those with reported expression in 0 or 1 of our tissues.

For the analysis of local similarity within different breadth classes, we classified genes according to the number of tissues with expression reported in human SAGE experiments: 0–1 (tissue-specific), 2–4, 5–8, 9–13, and 14–19 (broadly expressed).

Focal Average Correlation (ρ) and Statistics

In a modification of the method of Lercher et al. (2001), we first calculated focal averages of substitution rates. For each gene, we identified all other genes within 1 Mb along the chromosome, and calculated their mean substitution rate. On average, each gene had eight and nine such neighbors in the human–mouse and mouse–rat comparisons, respectively. When including only nondisparately evolving gene pairs, this was reduced to four and eight genes, respectively. When restricting the analysis to a certain class of genes, we included only focal genes and only neighboring genes within the same class. We then calculated Pearson's correlation coefficient across all data pairs consisting of the rate of a gene and the corresponding focal average; we denote this focal average correlation by ρ. A randomization protocol was employed to assess statistical significance. Gene positions were randomly permuted $N_0 = 10{,}000$ times, and ρ_{rand} was calculated for each random data set. n_P was the number of random data sets for which $\rho_{rand} \geq \rho$. From this, we estimated $P = (n_P + 1) / (N_0 + 1)$. This protocol maintains the original data structure; in particular, it leaves the distribution of neighbors unchanged.

To assess the correlation of K_i or K_{indel} values across different introns of the same gene, we similarly chose a focal intron and defined the focal average as the mean over all other introns in that gene. We then calculated Spearman's correlation coefficient (ρ) for data pairs consisting of the focal introns and their focal average. Statistical significance was estimated as above.

Throughout the paper, r denotes Pearson's correlation coefficient. Statistical significance was estimated by randomly permuting the values in one of the two columns. Repeating this $N_0 = 100{,}000$ times, we counted the number of times n_P when $r_{rand}^2 \geq r^2$, where r_{rand} is the correlation coefficient for the randomized data. From this, we estimate a two-sided $P = (n_P + 1) / (N_0 + 1)$. All correlation and focal average analyzes were also performed for Spearman's rank correlation coefficient, with very similar results (data not shown).

Before calculating linear regressions, we log-transformed values for substitution rates and GC measures.

ACKNOWLEDGMENTS

We thank Laurent Duret and Itai Yanai for helpful discussions, and four anonymous reviewers for comments on the manuscript. M.J.L. acknowledges financial support by The Wellcome Trust and the Royal Society. J.V.C. and L.D.H. are funded by the UK Biotechnology and Biological Sciences Research Council.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Aguilera, A. 2002. The connection between transcription and genomic instability. *EMBO J.* **21:** 195–201.

Arndt, P.F., Petrov, D.A., and Hwa, T. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol. Biol. Evol.* **20:** 1887–1896.

Averof, M., Rokas, A., Wolfe, K.H., and Sharp, P.M. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287:** 1283–1286.

Bielawski, J.P., Dunn, K.A., and Yang, Z. 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution: Approximate and maximum-likelihood methods lead to different conclusions. *Genetics* **156:** 1299–1308.

Bierne, N. and Eyre-Walker, A. 2003. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates. Implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics* **165:** 1587–1597.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., et al. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289–1292.

Casane, D., Boissinot, S., Chang, B.H.J., Shimmin, L.C., and Li, W.H. 1997. Mutation pattern variation among regions of the primate genome. *J. Mol. Evol.* **45:** 216–226.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17:** 540–552.

———. 2002a. Estimation of genetic distances from human and mouse introns. *Genome Biol.* **3:** RESEARCH0028.

———. 2002b. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. *Nucleic Acids Res.* **30:** 1751–1756.

Chamary, J.V. and Hurst, L.D. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: Evidence for selectively driven codon usage. *Mol. Biol. Evol.* (in press).

Datta, A. and Jinks-Robertson, S. 1995. Association of increased spontaneous mutation-rates with high-levels of transcription in yeast. *Science* **268:** 1616–1619.

Duret, L. and Mouchiroud, D. 2000. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17:** 68–74.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162:** 1837–1847.

Ebersberger, I., Metzler, D., Schwarz, C., and Paabo, S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70:** 1490–1497.

Eyre-Walker, A. 1993. Recombination and mammalian genome evolution. *Proc. R. Soc. London Ser. B* **252:** 237–243.

———. 1999. Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics* **152:** 675–683.

Filatov, D.A. 2003. A gradient of silent substitution rate in the human pseudoautosomal region. *Mol. Biol. Evol.* **21:** 410–417.

Filatov, D.A. and Gerrard, D.T. 2003. High mutation rates in human and ape pseudoautosomal genes. *Gene* **317:** 67–77.

Fullerton, S.M., Bernardo Carvalho, A., and Clark, A.G. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* **18:** 1139–1142.

Green, P., Ewing, B., Miller, W., Thomas, P.J., Nc, N., and Green, E.D. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33:** 514–517.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13:** 13–26.

Hastings, K.E.M. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. *J. Mol. Evol.* **42:** 631–640.

Hellmann, I., Ebersberger, I., Ptak, S.E., Paabo, S., and Przeworski, M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72:** 1527–1535.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.

Hughes, A.L. and Yeager, M. 1997. Comparative evolutionary rates of introns and exons in murine rodents. *J. Mol. Evol.* **45:** 125–130.

Huminiecki, L., Lloyd, A.T., and Wolfe, K.H. 2003. Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics* **4:** 31.

Hurst, L.D. and Eyre-Walker, A. 2000. Evolutionary genomics: Reading the bands. *Bioessays* **22:** 105–107.

Hurst, L.D. and Williams, E.J.B. 2000. Covariation of GC content and the silent site substitution rate in rodents: Implications for methodology and for the evolution of isochores. *Gene* **261:** 107–114.

Kondrashov, A.S. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mut.* **21:** 12–27.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.
Kumar, S. and Gadagkar, S.R. 2001. Disparity index: A simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics* **158:** 1321–1327.
Kumar, S. and Subramanian, S. 2002. Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci.* **99:** 803–808.
Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J., and Altschul, S.F. 2000. SAGEmap: A public gene expression resource. *Genome Res.* **10:** 1051–1060.
Lercher, M.J. and Hurst, L.D. 2002a. Can mutation or fixation biases explain the allele frequency distribution of human single nucleotide polymorphisms (SNPs)? *Gene* **300:** 53–58.
———. 2002b. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18:** 337–340.
Lercher, M.J., Williams, E.J.B., and Hurst, L.D. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human–rodent and mouse–rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* **18:** 2032–2039.
Lercher, M.J., Smith, N.G., Eyre-Walker, A., and Hurst, L.D. 2002a. The evolution of isochores: Evidence from SNP frequency distributions. *Genetics* **162:** 1805–1810.
Lercher, M.J., Urrutia, A.O., and Hurst, L.D. 2002b. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31:** 180–183.
Lercher, M.J., Urrutia, A.O., Pavlicek, A., and Hurst, L.D. 2003. A unification of mosaic structures in the human genome. *Hum. Mol. Genet* **12:** 2411–2415.
Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36:** 96–99.
———. 1997. *Molecular evolution.* Sinauer, Sunderland, Mass.
Majewski, J. and Ott, J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12:** 1827–1836.
Malcom, C.M., Wyckoff, G.J., and Lahn, B.T. 2003. Genic mutation rates in mammals: Local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol. Biol. Evol.* **20:** 1633–1641.
Margulies, E.H., Kardia, S.L., and Innis, J.W. 2001. Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res.* **29:** e60.
Matassi, G., Sharp, P.M., and Gautier, C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9:** 786–791.
Mellon, I., Spivak, G., and Hanawalt, P.C. 1987. Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian Dhfr gene. *Cell* **51:** 241–249.
Meunier, J. and Duret, L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* (in press).
Nachman, M.W. and Crowell, S.L. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156:** 297–304.
Nekrutenko, A. and Li, W.H. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.* **10:** 1986–1995.
Ogata, H., Fujibuchi, W., and Kanehisa, M. 1996. The size differences among mammalian introns are due to the accumulation of small deletions. *FEBS Lett.* **390:** 99–103.
Perry, J. and Ashworth, A. 1999. Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* **9:** 987–989.
Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., et al. 1996. A gene map of the human genome. *Science* **274:** 540–546.
Selby, C.P. and Sancar, A. 1993. Transcription-repair coupling and mutation frequency decline. *J. Bacteriol.* **175:** 7509–7514.
Silva, J.C. and Kondrashov, A.S. 2002. Patterns in spontaneous mutation revealed by human–baboon sequence comparison. *Trends Genet.* **18:** 544–547.
Smith, N.G.C. and Eyre-Walker, A. 2001. Synonymous codon bias is not caused by mutation bias in G+C- rich genes in humans. *Mol. Biol. Evol.* **18:** 982–986.
Smith, N.G.C. and Hurst, L.D. 1998. Sensitivity of patterns of molecular evolution to alterations in methodology: A critique of Hughes and Yeager. *J. Mol. Evol.* **47:** 493–500.
———. 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153:** 1395–1402.
Smith, N.G. and Lercher, M.J. 2002. Regional similarities in polymorphism in the human genome extend over many megabases. *Trends Genet.* **18:** 281–283.
Smith, N.G.C., Webster, M.T., and Ellegren, H. 2002. Deterministic mutation rate variation in the human genome. *Genome Res.* **12:** 1350–1356.
———. 2003. A low rate of simultaneous double-nucleotide mutations in primates. *Mol. Biol. Evol.* **20:** 47–53.
Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci.* **99:** 4465–4470.
Svejstrup, J.Q. 2002. Mechanisms of transcription-coupled DNA repair. *Nat. Rev. Mol. Cell Biol.* **3:** 21–29.
Tamura, K. and Kumar, S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol. Biol. Evol.* **19:** 1727–1736.
Tamura, K. and Nei, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10:** 512–526.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. Clustalw—Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Urrutia, A.O. and Hurst, L.D. 2003. The signature of selection mediated by expression on human genes. *Genome Res.* **13:** 2260–2264.
van Gool, A.J., van der Horst, G., Citterio, E., and Hoeijmakers, J.H.J. 1997. Cockayne syndrome: Defective repair of transcription? *EMBO J.* **16:** 4155–4162.
Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.
Versteeg, R., van Schaik, B.D.C., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J., and van Kampen, A.H.C. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13:** 1998–2004.
Williams, E.J.B. and Hurst, L.D. 2000. The proteins of linked genes evolve at similar rates. *Nature* **407:** 900–903.
———. 2002. Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. *J. Mol. Evol.* **54:** 511–518.
Yi, S.J., Ellsworth, D.L., and Li, W.H. 2002. Slow molecular clocks in Old World monkeys, apes, and humans. *Mol. Biol. Evol.* **19:** 2191–2198.

WEB SITE REFERENCES

http://www.ensembl.org; ENSEMBL Project home page.
http://genome.ucsc.edu; UCSC Genome Bioinformatics.
http://homepages.ed.ac.uk/eang33/mcinstructions.html; MCALIGN home page.
http://www1.imim.es/~castresa/Gblocks/Gblocks.html; GBLOCKS home page.
http://ncbi.nlm.nih.gov/dbEST; NCBI EST database home page.
ftp://ncbi.nlm.nih.gov/repository/UniGene; NCBI UniGene FTP site.
ftp://ncbi.nlm.nih.gov/pub/sage; NCBI SAGE FTP site.
ftp://ncbi.nlm.nih.gov/refseq; NCBI RefSeq FTP site.

Received May 28, 2003; accepted in revised form February 27, 2004.

SUPPLEMENTARY TABLES

Table S1. Benchmarking test of the predictive power of our estimate of housekeeping transcription rates. *r* is Pearson's correlation coefficient between the rate observed in the tissue listed in the left column, and the rate estimated *without* this tissue. In each case, we only included housekeeping genes. *r* was calculated including cases where observed transcription rate was zero in the excluded tissue.

Table S1.A. Correlation between predicted and observed human SAGE expression rate for housekeeping genes (prediction = median excluding the predicted tissue). All highly significant ($P<0.0001$, liver: $P=0.0056$)

tissue	*r*
mammary_gland	0.86
leukocyte	0.788
prostate	0.756
ovary	0.75
spinal_chord	0.748
lung	0.743
retina	0.738
stomach	0.738
vascular	0.728
kidney	0.705
heart	0.698
cortex	0.678
peritoneum	0.667
thalamus	0.66
kidney_embryonic	0.649
cerebellum	0.648
epidermis	0.642
astrocyte	0.53
liver	0.204

Table S1.B. Correlation between predicted and observed human microarray expression rate for housekeeping genes (prediction = median excluding the predicted tissue; anything below 0 was set to 0) (all highly significant, $P<0.0001$)

tissue	*r*
uterus	0.949
trachea	0.929
salivary_gland	0.912
thyroid	0.908
fetal_liver	0.900
spleen	0.895
whole_blood	0.894
pancreas	0.893
spinal_cord	0.887
placenta	0.885
ovary	0.884
cortex	0.871
amygdala	0.866
umbilical_vein	0.853
liver	0.843
heart	0.835
testis	0.834
cerebellum	0.833
thalamus	0.813
pituitary_gland	0.805
thymus	0.768
kidney	0.760
lung	0.760
caudate_nucleus	0.746
prostate	0.745
corpus_callosum	0.732
dorsal_root	0.730

Table S1.C. Correlation between predicted and observed mouse microarray expression rate for housekeeping genes (prediction = median excluding the predicted tissue; anything below 0 was set to 0) (all highly significant, $P<0.0001$)

tissue	*r*
frontal_cortex	0.922
spinal_cord_lower	0.897
eye	0.881
thyroid	0.877
mammary_gland	0.867
dorsal_root_ganglion	0.861
salivary_gland	0.850
lymph_node	0.844
adrenal_gland	0.841
uterus	0.836
stomach	0.831
placenta	0.829
olfactory_bulb	0.821
trigeminal	0.814
skeletal_muscle	0.805
tongue	0.789
bone	0.788
bone_marrow	0.770
hypothalamus	0.765
umbilical_cord	0.762
amygdala	0.752
spinal_cord_upper	0.751
ovary	0.748
gall_bladder	0.746
testis	0.746
kidney	0.745
brown_fat	0.737
digits	0.724
large_intestine	0.722
bladder	0.719
prostate	0.712
heart	0.710
spleen	0.710
hippocampus	0.702
cortex	0.694
snout_epidermis	0.675
trachea	0.673
epidermis	0.668
liver	0.665
striatum	0.632
cerebellum	0.610
thymus	0.600
small_intestine	0.558
lung	0.529

Table S1.D. Correlation between predicted and observed mouse SAGE expression rate for housekeeping genes (prediction = median of positive values, excluding the predicted tissue) (all highly significant, $P<0.0001$)

tissue	*r*
testis_fetal	0.618
nucleus_accumbens	0.568
cerebellum	0.561
cortex	0.541
granular_cell_precursor	0.531
striatum	0.496
forelimb_embryo	0.478
hindlimb_embryo	0.476
testis	0.377

Table S2. Difference of synonymous substitution rate excluding CpG sites (K_4^{excCpG}) between housekeeping and tissue specific genes

		human-mouse					mouse-rat				
	breadth measure	house-keeping[1]	tissue-specific[1]	ΔK_4[2]	P[3]	N	house-keeping[1]	tissue-specific[1]	ΔK_4[2]	P[3]	N
K_4											
	EST	0.29	0.368	0.078	0.0047	115	0.102	0.116	0.014	0.00016	453
	SAGE	0.282	0.349	0.067	0.00007	377	0.099	0.113	0.014	<0.00001	1901
	microarray	0.29	0.364	0.074	0.00016	99	0.104	0.11	0.006	0.026	777
K_4 residuals from regression on K_A											
	EST	-0.006	0.023	0.029	0.12	115	0.0071	0.0082	0.001	0.39	453
	SAGE	-0.004	0.003	0.007	0.34	377	0.0034	0.0077	0.004	0.055	1901
	microarray	0.004	0.020	0.016	0.20	99	0.0056	0.0047	-0.001	0.62	777

[1] Average K_4^{excCpG} or average residuals of K_4^{excCpG}. Residuals were calculated from expected K_4^{excCpG} values, which were predicted from linear regression of $\log(K_4^{excCpG})$ on $\log(K_A)$ including all genes. Genes were classified as housekeeping / tissue specific if supported by experiments in both human and mouse for the human-mouse comparison, and by experiments in mouse for the mouse-rat comparison. Only non-disparate gene pairs were included.

[2] Difference in K_4^{excCpG} (or residuals) between tissue specific and housekeeping averages.

[3] Probability of finding an equal or greater difference in 100,000 randomised genomes.

Table S3. Correlation between synonymous substitution rate excluding CpG sites (K_4^{excCpG}) and expression breadth

		human-mouse			mouse-rat		
	breadth measure[1]	$r^{(2)}$	P	N	$r^{(2)}$	P	N
all genes							
	EST	-0.142	<0.00001	3191	-0.071	0.00010	3110
	SAGE	-0.104	<0.00001	2731	-0.095	<0.00001	2913
	microarray	-0.16	<0.00001	1508	-0.065	0.0061	1779
GC≤0.5							
	EST	-0.261	<0.00001	976	-0.183	<0.00001	648
	SAGE	-0.151	0.00003	806	-0.221	<0.00001	641
	microarray	-0.256	<0.00001	423	-0.082	0.11	383

[1] Breadth of expression was averaged over experiments in human and mouse for the human-mouse comparison, and estimated from mouse data for the mouse-rat comparison.

[2] Pearson's correlation coefficient between expression breadth and K_4. Only non-disparate gene pairs were included.

Table S4. Effect of expression breadth on significance of local similarity (ρ) in K_4^{excCpG}

	human-mouse				mouse-rat			
breadth measure[1]	ρ	$P_{all}^{(2)}$	$P_{group}^{(3)}$	N	ρ	$P_{all}^{(2)}$	$P_{group}^{(3)}$	N
EST	0.220	<0.0001	<0.0001	889	0.121	0.0003	<0.0001	2487
SAGE	0.250	<0.0001	<0.0001	700	0.155	0.0003	<0.0001	2303
microarray	0.356	<0.0001	<0.0001	304	0.091	0.001	0.0008	1293

[1] Breadth of expression was averaged over experiments in human and mouse for the human-mouse comparison, and is obtained from mouse only in the mouse-rat comparison. Only non-disparate gene pairs were included.

[2] P_{all} is the fraction of equal or greater ρ in datasets obtained by randomly permuting all genes

[3] P_{group} is the fraction of equal or greater ρ in datasets obtained by randomly permuting genes within classes of similar K_4

Table S5. Correlation between expression rate and K_4^{excCpG} for putative housekeeping genes

	human-mouse			mouse-rat		
rate measure[1]	*r*	*P*	*N*	*r*	*P*	*N*
non-disparate						
SAGE (human)	-0.028	0.752	134	-	-	-
microarray (human)	-0.139	0.077	163	-	-	-
SAGE (mouse)	-0.103	0.202	156	0.0266	0.65	299
microarray (mouse)	-0.081	0.271	186	-0.0344	0.51	382
non-disparate & GC≤0.5						
SAGE (human)	-0.474	0.003	37	-	-	-
microarray (human)	-0.117	0.532	31	-	-	-
SAGE (mouse)	-0.371	0.012	45	0.0042	0.97	80
microarray (mouse)	-0.161	0.254	52	0.036	0.73	93

[1] Rate measures were not averaged over human and mouse in this table to retain acceptable sample sizes. Only non-disparate gene pairs were included.

Table S6. Effect of K_4^{excCpG} on significance of local similarity in K_A (duplicate genes excluded)

	human-mouse				mouse-rat			
	ρ	P_{all}[1]	P_{group}[(2)]	*N*	ρ	P_{all}[1]	P_{group}[(2)]	*N*
all genes	0.053	0.0013	0.12	4087	0.026	0.061	0.074	3787
non-disparate	0.068	0.016	0.23	1224	0.039	0.019	0.019	3264

[1] P_{all} is the number of equal or greater ρ in datasets obtained by randomly permuting all genes

[2] P_{group} is the number of equal or greater ρ in datasets obtained by randomly permuting genes within classes of similar K_4

Part II. **Selection at silent sites in introns and exons**

Regional variation in rates of evolution (Part I) occurs at the scale of megabases, and so cannot be explained by differences between autosomes. If silent sites evolve neutrally, this means that I effectively show that the mutation rate is heterogeneous across autosomes, and hence between genes. But do silent rates of evolution necessarily reflect the mutation rate? Chapter 3 is an indirect test for neutrality, whereby one compares two groups of silent sites and asks whether they differ.

Previously, several groups have found that pseudogenes evolve faster than synonymous sites in their functional counterparts, which suggests the action of purifying selection at synonymous sites (Miyata & Hayashida 1981; Bustamante, Nielsen & Hartl 2002). Unfortunately, however, this suffers from several confounding factors that render interpretation difficult. First, only transcribed genes will experience biases associated with transcriptional-coupled mutation and repair (Green et al. 2003; Majewski 2003). Second, the two groups of sequence may be present in different isochores. Differences in GC content will, for example, affect the relative abundance of hypermutable CpG dinucleotides. Third, as I described in Part I, evolutionary rates vary across the genome.

Consequently, it is desirable to carry out pairwise tests within the same gene. Iida and Akashi (2000), for example, compared constitutively and alternatively expressed exons. They hypothesised that, because constitutive exons are translated more frequently than alternative exons, a difference in nucleotide content, reflecting the use of optimal codons, would indicate selection. Indeed, they found in mammals that both GC3 (GC content at the mostly synonymous codon third sites) and the rate of synonymous evolution are higher in exons that are expressed constitutively.

In Chapter 3, I compare patterns of evolution at two classes of silent sites within murid genes, in introns and at four-fold degenerate (synonymous) sites. Before making the comparison, I first ask whether there is putatively neutral sequence that we expect *a priori* to be conserved. Hence the chapter is divided into two parts.

Although there are plentiful descriptions in the molecular biology literature of transcriptional control elements within first introns, only a couple have compared their activity relative to the other introns within a gene (Palmiter et al. 1991; Jonsson et al. 1992). As these anecdotes might reflect a reporting bias, I asked whether it was generally true that first introns contain disproportionately more control elements and evolve more slowly as a result. While I found no difference between the densities of transcription factor binding sites, first introns do possess a higher density of CpG

islands. Although several authors have shown that first introns evolve slowly (Majewski & Ott 2002; Keightley & Gaffney 2003), none have systematically shown that this is because they contain more control elements. I show that substitution rates are lower in first introns and that this is consistent with purifying selection to preserve the activity of control elements. Additionally, I find that substitutions are increasingly less frequent as one approaches the intron-exon junction, which I assume reflects selection for splice site recognition.

In the second part of Chapter 3, after eliminating selectively constrained intronic sites, I compare the rates of substitution between introns and four-fold sites. Consistent with the predictions of the neutral theory, these two classes of putatively neutral sites evolve at equivalent rates. Although at first sight this does not deviate from the null expectation, neutrality also predicts that the two classes of sites not only evolve at the same rate, but also in the same manner. Upon closer inspection, however, this is not observed. By examining the patterns of substitution at the first site within dinucleotides, it can be seen that As and Ts are rarely conserved, but C is particularly stable given its relative abundance. Similarly, while A and T content are both higher in introns than at four-fold sites, G content is the same while C content is higher at four-fold sites. Overall, if one assumes that the majority of the remaining intronic sequence is less likely to be under selective constraint, these results suggests that the C preference at four-fold sites provides evidence for selection at synonymous sites.

References

Bustamante, C. D., Nielsen, R. & Hartl, D. L. (2002) A maximum likelihood method for analyzing pseudogene evolution: Implications for silent site evolution in humans and rodents. *Mol. Biol. Evol.* **19**: 110-117.

Green, P., Ewing, B., Miller, W., Thomas, P. J., Nc, N. & Green, E. D. (2003) Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33**: 514-517.

Iida, K., & Akashi, H. (2000) A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene* **261**: 93-105.

Jonsson, J. J., Foresman, M. D., Wilson, N. & McIvor, R. S. (1992) Intron requirement for expression of the human purine nucleoside phosphorylase gene. *Nucleic Acids Res.* **20**: 3191-3198.

Keightley, P. D., & Gaffney, D. J. (2003) Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl Acad. Sci. USA* **100**: 13402-13406.

Majewski, J., & Ott, J. (2002) Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12**: 1827-1836.

Majewski, J. (2003) Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am. J. Hum. Genet.* **73**: 688-692.

Miyata, T., & Hayashida, H. (1981) Extraordinarily high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. *Proc. Natl Acad. Sci. USA* **78**: 5739-5743.

Palmiter, R. D., Sandgren, E. P., Avarbock, M. R., Allen, D. D. & Brinster, R. L. (1991) Heterologous Introns Can Enhance Expression of Transgenes in Mice. *Proc. Natl Acad. Sci. USA* **88**: 478-482.

Chapter 3. **Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage**

Jean-Vincent Chamary & Laurence D. Hurst

Molecular Biology and Evolution (2004) **21**: 1014-1023

Similar Rates but Different Modes of Sequence Evolution in Introns and at Exonic Silent Sites in Rodents: Evidence for Selectively Driven Codon Usage

Jean-Vincent Chamary and Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

In mammals divergence at fourfold degenerate sites in codons (K_4) and intronic sequence (K_i) are both used to estimate the mutation rate, under the supposition that both evolve neutrally. Does it matter which of these we use? Using either class of sequence can be defended because (1) K_4 is the same as K_i (at least in rodents) and (2) there is no selectively driven codon usage (hence no systematic selection on third sites). Here we re-examine these findings using 560 introns (for 136 genes) in the mouse-rat comparison, aligned by eye and using a new maximum likelihood protocol. We find that the rate of evolution at fourfold sites and at intronic sites is similar in magnitude, but only after eliminating putatively constrained sites from introns (first introns and sites flanking intron-exon junctions). Any approximate congruence between the two rates is not, however, owing to an underlying similarity in the mode of sequence evolution. Some dinucleotides are hypermutable and differently abundant in exons and introns (e.g., CpGs). More importantly, after controlling for relative abundance, all dinucleotides starting with A or T are more prevalent in mismatches in exons than in introns, whereas C-starting dinucleotides (except CG) are more common in introns. Although C content at intronic sites is lower than at flanking fourfold sites, G content is similar, demonstrating that there exists a strong strand-specific preference for C nucleotides that is unique to exons. Transcription-coupled mutational processes and biased gene conversion cannot explain this, as they should affect introns and flanking exons equally. Therefore, by elimination, we propose this to be strong evidence for selectively driven codon usage in mammals.

Introduction

In mammals, divergences of two classes of sequence are regularly used to estimate the mutation rate: fourfold degenerate sites in exons (Eyre-Walker and Keightley 1999; Keightley and Eyre-Walker 2000) and intronic DNA (Chang et al. 1994; Chang and Li 1995; Chang, Hewett-Emmett, and Li 1996; Huang et al. 1997). That the rates of point substitution at both classes of site (K_4 and K_i, respectively) are valid measures is supported by two important findings. First, there is the finding that in rodents K_4 (or K_s) and K_i are approximately equal (Hughes and Yeager 1997), suggesting that the mode of evolution (putatively neutral) is the same in the two classes of sequence. Second, unlike most taxa (e.g., bacteria, yeast, fly, and nematode), there is no selectively driven codon usage in mammalian genes (Eyre-Walker 1991; Smith and Hurst 1999*a*; Kanaya et al. 2001; Duret 2002). This is evidenced by (among other things) the lack of correspondence between the usage of a codon and iso-acceptor tRNA abundance (Duret 2002).

Both of these findings require re-analysis, particularly because accumulating evidence suggests that neither fourfold degenerate sites nor introns are entirely free of constraint. Hughes and Yeager (1997) used complete intron sequences and all introns from a gene (except those too difficult to align). While noting that the splicing control regions that flank intron-exon junctions are subject to selective constraint, they reasonably argued that the number of such sites is too small to matter. However, recent evidence suggests that sequence conservation associated with splice sites may extend relatively far away from intron-exon boundaries (Majewski and Ott 2002; Waterston et al. 2002; Hare and Palumbi 2003; Sorek and Ast 2003). Majewski and Ott (2003), for example, showed that human SNP density and SINE insertion frequency is lower in the first and last 20 bp of introns and constraint may extend up to 200 bp into intronic sequence. The extent of conservation may well differ between the 5′ and 3′ ends (Majewski and Ott 2002; Sorek and Ast 2003). Given uncertainty over the size of the conserved region, we start by estimating its average size. We then purge our intronic alignments of these regions.

Mammalian introns also contain other motifs that could be under purifying selection, such as transcription factor (TF) binding sites (e.g., Rossi and de Crombrugghe 1987; Katai et al. 1992; Kawada et al. 1999; Suen and Goss 2001). The presence of such control elements may explain why transgene expression can be 10–100 times more efficient when introns are added to cDNA clones (Brinster et al. 1988). An estimated 10% of mouse introns contain regulatory elements, of which a fraction overlaps with CpG islands (Waterston et al. 2002).

Intron-associated regulatory elements are believed to be nonrandomly distributed within a gene. For example, they tend to be located in close proximity to the start codon, and thus in the first intron within the coding sequence (Sakurai et al. 2002). This in turn may explain (Sakurai et al. 2002) why the intron in single-intron genes tends to be located 5′ end (see also Fink 1987; Mourier and Jeffares 2003). Numerous reports (e.g., Oshima, Abrams, and Kulesh 1990; Rohrer and Conley 1998; Chan et al. 1999) describe control elements in first introns (see also all the above references describing introns with TF binding sites). Although only a few studies have compared all introns derived from the same gene (e.g., Palmiter et al. 1991; Jonsson et al. 1992), these report that the first intron has the greatest impact in modulating expression. In contrast, a systematic in silico analysis on a large data set failed to identify a higher frequency of TF binding sites in first introns (Levy,

Key words: codon usage bias, point substitution rate, purifying selection, introns, fourfold degenerate sites, dinucleotides.

E-mail: l.d.hurst@bath.ac.uk.

Mol. Biol. Evol. 21(6):1014–1023. 2004
DOI:10.1093/molbev/msh087
Advance Access publication March 10, 2004

Hannenhalli, and Workman 2001). Although this may reflect our poor abilities to detect control elements computationally, some control elements have been described in non-first introns (e.g., Lothian and Lendahl 1997; Hural et al. 2000). The 5′ end of first introns may be of particular importance in transcriptional control (Majewski and Ott 2002).

If it is a general property of first introns to harbour more control elements, we would expect them to evolve slower than the other "non-first" introns, all things being equal. However, as first introns tend to be larger (Hawkins 1988; Smith 1988), a higher number of constrained sites need not imply a higher density of constrained sites. Indeed, Levy, Hannenhalli, and Workman (2001) report that, if anything, first introns evolve faster. However, this analysis came from the mouse-human comparison in which alignment of freely evolving sites is unreliable (Jareborg, Birney, and Durbin 1999). In contrast, we find that first introns evolve slower and that the reduced point substitution rate is in part owing to a greater abundance of CpG islands. Given this evidence for the likely action of purifying selection on intronic sites, we then ask whether K_4 becomes significantly lower than K_i after removing constrained sites. Such a finding could undermine the use of K_4 as an estimator of the mutation rate.

The second issue we investigate is whether selection on synonymous sites exists and, more specifically, whether we can detect selectively driven codon usage. Recent direct evidence shows that synonymous mutations can be highly deleterious (Duan et al. 2003). Further, codon usage bias has recently been reported (Urrutia and Hurst 2003) to be greater in highly expressed genes (see also Debry and Marzluff 1994). Comparably, constitutively expressed exons have a higher GC content than those that are alternatively expressed (Iida and Akashi 2000). It is possible that as many as 40% of fourfold degenerate sites are under selection (Hellmann et al. 2003).

To address this issue, here we ask whether the substitution processes in introns and at synonymous sites are comparable. Given that certain dinucleotides can be differently abundant in exons and introns (e.g., "differential CpG content," Subramanian and Kumar 2003; see also Hellmann et al. 2003), we analyze all possible dinucleotides and ask whether there is a discrepancy in their relative abundances between fourfold degenerate sites and introns. We then determine the "mutability" of each dinucleotide, i.e., how often a given dinucleotide is associated with a mismatch in introns and at fourfold sites. Additionally, a dinucleotide could be equally abundant in the two classes of sequence, but have different forces operating on it. To examine this we investigate the "stability" of each dinucleotide, i.e., the rate of involvement of a given dinucleotide in a mismatch, after controlling for the abundance of the dinucleotide.

Materials and Methods

Data Set of Orthologous Mouse-Rat Genes

We expanded upon a data set of over 40 mouse-rat orthologs (Hughes and Yeager 1997; Smith and Hurst 1998), where orthology was determined through HOVERGEN (the Homologous Vertebrate Gene Database, Release 42, available at www.hgmp.mrc.ac.uk; Duret, Mouchiroud, and Gouy 1994). Each gene pair is considered orthologous only if, within the gene family tree, there is no non-rodent lineage between the mouse and rat branches and if at least one non-rodent sequence is present as an outgroup. Orthology was further validated through syntenic comparisons using LocusLink (www.ncbi.nlm.nih.gov/LocusLink/) and the Rat Genome Database (RGD) Virtual Comparative Map tool (http://rgd.mcw.edu/VCMAP/).

A list of 5,339 rat genes was downloaded from the RGD (http://rgd.mcw.edu/pub/data_release/GENES). The corresponding sequence entries were extracted from GenBank (www.ncbi.nih.gov/Genbank). Each GenBank file was scrutinized for the presence of annotations describing the location of every exon. This returned 231 matches for complete genes possessing at least one intron. Excluding the rat genes for which a confirmed mouse ortholog had already been identified, each of the remaining 189 rat sequences was Blasted (Altschul et al. 1990) against the complete mouse genome at Ensembl (version 14.30.1, www.ensembl.org/Mus_musculus/blastview/), returning 126 hits.

We examined each rat gene in HOVERGEN to identify the mouse ortholog. If the mouse ortholog had the introns described, the HOVERGEN-derived GenBank files replaced Blast ones (which occurred in 44 cases). If the HOVERGEN-described mouse ortholog was the same as the Blast return, but lacking in introns (i.e., only mRNA described), then we retained the Blast entry. In 23 cases, HOVERGEN described an unambiguous ortholog, different from the Blast match. These were eliminated from our data set. For the remaining 58 genes for which HOVERGEN did not specify a mouse ortholog (usually because no rat sequence was available for the gene family), the orthology of the Blast sequence to the rat sequence was confirmed by ensuring that the genes were syntenic with the orthologous region in rat, that intron pairs were well-aligned overall, and that the estimated rate of protein evolution was within normal bounds for the mouse-rat comparison ($K_a < 0.2$).

Of the remaining data set of 142 genes, a further 2 were excluded due to poor sequence annotation and another 4 because they were located on the X-chromosome (these having lower substitution rates than autosomal genes [Hurst and Ellegren 1998]). Syntenic comparisons and the RGD list of accession numbers for each gene were used to ensure that there was no redundancy within the data set. For statistical analyses, we report sample size as "N_g" if evaluated on a gene-by-gene basis and as "N_i" on an intron-by-intron basis.

Sequence Alignments

Coding sequence was extracted from GenBank files using GBPARSE (http://sunflower.bio.indiana.edu/~wfischer/Perl_Scripts/). Alignment of the translated sequences was carried out using PILEUP. Nucleotide alignments were reconstructed from the amino acid

sequence alignments using AA2NUC (available from L.D.H.).

Only internal introns located between coding exons were analyzed. Two methods were used for aligning introns: (1) manually (by-eye) and (2) using MCALIGN, a stochastic maximum likelihood (ML)–based program incorporating a Monte Carlo algorithm (P. D. Keightley and T. Johnson, unpublished data, http://homepages.ed.ac.uk/eang33/mcinstructions.html). MCALIGN is based on a model of noncoding sequence evolution that is built upon the frequency of indel events relative to nucleotide substitutions. We executed the program using the rodent intron parameters provided. Seven massive introns (>7 kb, two of which were first introns) proved too difficult to align. Our final data set consisted of 136 orthologous genes possessing 560 introns.

Estimating Rates of Evolution

Both K_4, the number of substitutions per fourfold degenerate site within exons, and K_i, the intronic substitution rate, were estimated using the algorithmic method of Tamura and Nei (1993). To obtain genic K_i values, intronic substitution rates were weighted according to the number of bases compared per individual intron alignment (Smith and Hurst 1998). The indel rate, K_{indel}, was calculated as the total number of indels per base pair of the alignment.

Accounting for Alignment Artifact

After estimating the K_i per intron by the two alignment methods, we defined a conservative set and a liberal set. The former contains the lower estimate of K_i, and the latter contains the higher value. For the intron alignments in the conservative alignment set, 516 introns were drawn from the by-eye set and the remainder from the ML set. Both alignment methods agree exactly on both K_{indel} and K_i for 84 of our orthologous introns. We define this as our "tight" set ($N_g = 50$). This subset consists mainly of short introns (mean length 122 bp, as opposed to 604 bp).

Alignment-induced noise is commonly minimized by rejecting difficult-to-align introns from analysis (e.g., Hughes and Yeager 1997; Smith and Hurst 1998). Instead, we filtered out ambiguous regions of alignments to produce a fifth alignment set containing all introns ($N_i = 560$). We did this by applying the Gblocks program (www1.imim.es/~castresa/Gblocks/Gblocks.html; Castresana 2000) to the alignments in the conservative set under the default parameters.

Sequence Conservation at Intron-Exon Junctions

For a given distance away from the intron-exon boundary (running 5′ → 3′ for the 5′ end, 3′ → 5′ for the 3′ end), we calculated the frequency among all introns of mismatches at the site in question. Mismatches were defined in two ways: (1) a nucleotide aligned against a different nucleotide (i.e., those that contribute to estimates of K_i) and (2) a nucleotide matched with a gap. Ambiguous nucleotides (N) were ignored.

Detecting Control Elements and Transcription Factor Binding Sites

To ask whether first introns contain more regulatory elements, and whether the presence of regulatory elements predicts the rate of evolution, we assayed introns for the presence of CpG islands and transcription factor (TF) binding sites.

CpG islands tend to be located in regions with higher than expected numbers of CpG dinucleotides in the coding strand (Gardiner-Garden and Frommer 1987). We used the recommended settings for detecting CpG islands at the 5′ of genes implemented by CPGREPORT (available as part of EMBOSS, www.ebi.ac.uk/emboss/cpgplot/), which carries out a running-sum (not window) analysis to identify CpGs associated with putative islands. CpGs in putative islands are not necessarily present as a continuous array, so we also define islands by the extent to which the CpGs in such islands are clustered within a CpG-rich region. We therefore define island density by whether we require a minimum of 0, 1, 5, 10, 15, or 20 clustered island-associated CpGs to define an island. The density of putative CpG islands within a given intron was taken as the mean of the densities in mouse and rat.

To identify TF binding sites, we scanned our sequences for exact matches to well-characterized vertebrate TF binding sites as described in TRANSFAC (Wingender et al. 2000). We did this by employing TFSCAN within EMBOSS (www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/tfscan.html). We then masked the corresponding input sequences at the positions where hits were found. We then determined, for each intron, the density of putative TF binding sites. We also examined mouse TF binding sites alone. While this qualitatively affects the observed densities, it does not affect any of the patterns that we observe. We report only the data from the vertebrate collection.

Dinucleotide Content and Mismatch Rates

Any given mismatch in an alignment is associated with four dinucleotides (e.g., ATT/ACT is associated with AT, AC, TT, and CT). At fourfold sites, however, there can be no dinucleotides starting with A that have the mismatch at the second site, as there are no fourfold degenerate codons with A at their second site. Therefore, to give a fairer impression of the forces acting in exons and introns, in exons we only consider dinucleotides in which the first base is the one at the fourfold site. To ensure comparability, in introns we only count dinucleotides as they occur at the first site of the pair. For each dinucleotide we calculate, for fourfold degenerate sites and introns: (1) the relative abundance (frequency of occurrence) of the dinucleotide; (2) "mutability," the probability with which the dinucleotide is associated with a mismatch; and (3) "stability," the probability with

Table 1
Estimates for Rates of Evolution from Mouse-Rat Orthologous Introns and Exons

	Alignment Set[b]					
Rate[a]	By-Eye	Max. Likelihood	Conservative	Liberal	Tight	Gblocks
K_i	0.1478 ± 0.0282	0.1701 ± 0.0356	0.1468 ± 0.0274	0.1710 ± 0.0358	0.1454 ± 0.0495	0.1443 ± 0.0266
K_i first	0.1442 ± 0.0399	0.1616 ± 0.0513	0.1431 ± 0.0394	0.1627 ± 0.0513	0.1375 ± 0.0448	0.1405 ± 0.0391
K_i non-first	0.1533 ± 0.0293	0.1791 ± 0.0406	0.1526 ± 0.0286	0.1798 ± 0.0408	0.1466 ± 0.0518	0.1504 ± 0.0271
K_a	0.0421 ± 0.0391					
K_s	0.1718 ± 0.0513					
K_4	0.1824 ± 0.0723					

[a] Mean point substitution rate (± SD). For a given gene, the intronic substitution rate (K_i) is the mean across all introns weighted by the size (number of aligned sites) of each intron. Exonic substitution rates (K_a, K_s, and K_4) are shown for estimates from all 136 genes.

[b] The number of genes (number of introns), N_g (N_i), used to calculate K_i (excluding the tight alignment set) = including all introns 136 (560), first 134 (134), non-first 118 (426). For the tight set, N_g (N_i) = all introns 50 (84), first introns 16 (16), non-first introns 42 (68).

which the dinucleotide is associated with a mismatch, per occurrence of the dinucleotide.

Results

Comparing Estimates of Evolutionary Rates Generated by Alternative Alignment Methods

How confident can we be that our estimates of intronic point substitution rates are accurate and relatively unaffected by alignment artifacts? We attempted to resolve this problem by aligning each orthologous intron pair using two different methods. Although there was a strong correlation between the values obtained per intron ($R^2 = 0.678$, $P < 0.001$, $N_i = 560$), the maximum likelihood (ML) program tends to produce slightly higher estimates for K_i than the by-eye method (table 1). For the 84 introns in the tight alignment set, K_i is very similar to that in the conservative alignment set. Thus, these are likely to be minimum estimates, whereas the liberal set reflects realistic upper estimates.

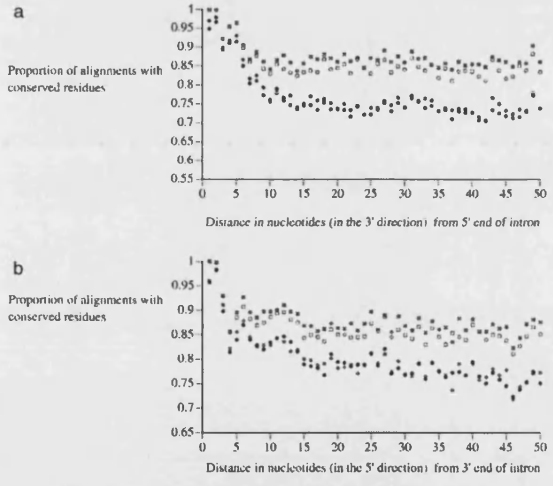


FIG. 1.—Sequence conservation at intronic sites flanking intron-exon junctions as a function of distance from the junction at the (*A*) 5′ end and (*B*) 3′ end. Conservation is defined with (circles) and without (squares) counting gaps as informative sites. The alignment methods are by-eye (grey) and using a maximum likelihood protocol (black/white).

Constraint on Intronic Sites: How Much Sequence Flanking Intron-Exon Junctions Is Conserved?

Considering figure 1, we suggest that, to be conservative, intron sequence within the first and last 20 bp should be excluded from analysis. As expected (Reed and Maniatis 1985), there is selection against substitutions at the first and last two intronic sites adjacent to splice junctions (fig. 1). In contrast, at the 3′ end, we do not find evidence that the 7-nucleotide branch site, commonly located 18–40 nucleotides upstream (Reed and Maniatis 1985), is located at any single well-conserved region. However, though the branch site is the preferred site for lariat formation, it is not essential (Zhuang, Goldstein, and Weiner 1989). On the other hand, we observe minor peaks in conservation approximately 12 bp and 19 bp upstream of the 3′ end, which may represent two alternative locations for the branch site (fig. 1*B*). As there is not enough evidence to justify exclusion of further nucleotides (e.g., up to 200 bp), we quote substitution rates for all analyses given below, after excluding only the 20 bp at each end.

Constraint on Intronic Sites: First Introns Contain More CpG Islands and Evolve Slower

Previous reports have shown that first introns can enhance gene expression to a greater degree than other introns from the same gene. Does this mean that first introns evolve slower? As we report in table 2, we find this to be the case (see also table 1). The obvious explanation is that first introns possess more control elements. The possession of CpG islands could have a twofold effect in slowing first intron evolution, not only by imposing functional constraints, but also by demethylation of CpG dinucleotides that would otherwise be hypermutable. We observe that first introns are richer in CpGs belonging to putative CpG islands (table 3). In contrast, we find no evidence that first introns have a higher density of transcription factor binding sites (see table *A* in the Supplementary Material online). The latter is a weak test, particularly because the short length of TF binding sites results in high rates of degeneration and spontaneous emergence ("turnover") (Dermitzakis and Clark 2002). Indeed, 30%–40% of binding sites known to be present in humans cannot be detected in rodents (Dermitzakis and

Table 2
Differences in Rates of Evolution and GC Content Between First and Non-first Introns from the Same Gene (One-Sample Wilcoxon Signed-Rank Tests, $N_g = 116$)

	Alignment Set								
	Conservative			Liberal			Gblocks		
Rate/GC Content	*P*-value	Mean first	Mean Non-first	*P*-value	Mean First	Mean Non-first	*P*-value	Mean First	Mean Non-first
K_i	0.031	0.1452 ± 0.0397	0.1515 ± 0.0283	0.002	0.1648 ± 0.0514	0.1783 ± 0.0402	0.029	0.1432 ± 0.0398	0.1494 ± 0.0268
K_i non-CpG	0.026	0.1274 ± 0.0403	0.1351 ± 0.0284	0.001	0.1457 ± 0.0521	0.1598 ± 0.045	<0.001	0.1176 ± 0.0395	0.1398 ± 0.0510
GCi	0.765	50.692 ± 7.936	50.483 ± 8.055						

Clark 2002). In mouse, the proportion of G + C nucleotides within regulatory elements is generally higher than overall genomic GC content (Waterston et al. 2002). However, we do not find a significant difference between mean GCi of first introns and non-first introns (table 2).

We expect that introns with control elements should evolve slower than those without. At least for putative CpG islands, this is so (table *B* in the Supplementary Material online). Could this, along with the greater abundance of CpG islands in first introns, entirely explain why first introns evolve slowly? Do first introns without CpG islands then have the same rate of evolution as non-first introns also lacking such islands? First introns still evolve more slowly (online supplementary table *C*), suggesting that CpG island presence only partly explains the difference between first and non-first introns. We find no correlation between putative TF binding site density and K_i (online supplementary table *D*).

We presumed that the excess of CpGs found in first introns reflects unmethylated CpG islands. The higher density of CpG-rich regions in first introns might instead represent methylated hypermutable CpGs, rather than conservation of control elements (CpG islands). To examine this we asked whether, after masking CpG dinucleotides in all introns, first introns have a lower rate of evolution than non-first. We find that the significance of the difference in K_i increases after masking (table 2), indicating that the CpGs present are constrained putative islands, not hypermutable sites.

Given that in silico methods have a high false positive rate (Fickett and Hatzigeorgiou 1997; Wasserman et al. 2000), can we be confident that removal of first introns also eliminates most control elements? We address this issue by asking whether, before and after removal of first introns, there is heterogeneity between introns in their rate of evolution. We classified introns according to their relative position within a given gene, i.e., 1 = first intron, 2 = second, etc. As expected, if all introns are analyzed, we observe highly significant heterogeneity in K_i between introns at different positions (e.g., liberal set: $P = 0.0019$, Kruskal-Wallis, df = 16). Importantly, after removal of first introns this heterogeneity disappears ($P = 0.1671$, df = 15). Similarly, there exists a highly significant correlation between intron position and K_i when first introns are included ($\rho^2 = 0.0278$, $P < 0.0001$, Spearman rank, $N_i = 560$) but not after their removal ($\rho^2 = 0.0037$, $P = 0.212$, $N_i = 426$). These findings are consistent with the notion that most functionally constrained elements are located in first introns.

Exclusion of Constrained Sites Leads to a Similar K_4 and K_i

If we include first introns in our estimate of K_i, we find that fourfold sites evolve faster than intronic ones (independent of the alignment set used, $P < 0.05$, one-sample Wilcoxon signed-rank tests, $N_g = 136$, tight set $N_g = 50$). In this regard, we fail to replicate the prior results (Hughes and Yeager 1997; Smith and Hurst 1998). However, if we exclude first introns, we find that $K_4 = K_i$ in the ML and liberal sets ($P > 0.3$, one-sample Wilcoxon signed-rank tests, $N_g = 118$) but not in the other sets ($P < 0.001$, one-sample Wilcoxon signed-rank tests, $N_g = 118$). We can conclude that fourfold sites do not evolve slower than intronic sites, and that after putatively constrained sites are removed from introns, K_4 and K_i become more similar in magnitude.

Association of Dinucleotides with Mismatches

Given that the rate of sequence evolution in introns (excluding sites under purifying selection) and at fourfold sites are about the same, should we conclude that the

Table 3
Differences in CpG Island Density Between First and Non-first Introns from the Same Gene (Paired *t*-tests, $N_g = 116$)

	Minimum Number of Clustered Island-Associated CpGs Required to Define a Putative CpG Island[a]					
Density	0	1	5	10	15	20
P-value	0.005	0.006	0.004	0.004	0.039	0.077
Ratio	1.3876	1.6459	3.3275	6.3027	9.4976	14.3903
Mean first	0.0381 ± 0.0407	0.0284 ± 0.0426	0.0167 ± 0.0424	0.0114 ± 0.0366	0.0063 ± 0.0261	0.0047 ± 0.0231
Mean non-first	0.0275 ± 0.0260	0.0173 ± 0.0263	0.0050 ± 0.0214	0.0018 ± 0.0125	0.0007 ± 0.0067	0.0003 ± 0.0038

[a] Putative CpG islands become increasingly more difficult to detect as the threshold for the number of island-associated CpGs required to define an island increases, so when 20 clustered CpGs are required, the majority of introns have no detectable CpG cluster and the significance level is dependent on the number ($N_i = 15$) in which the islands can be detected.

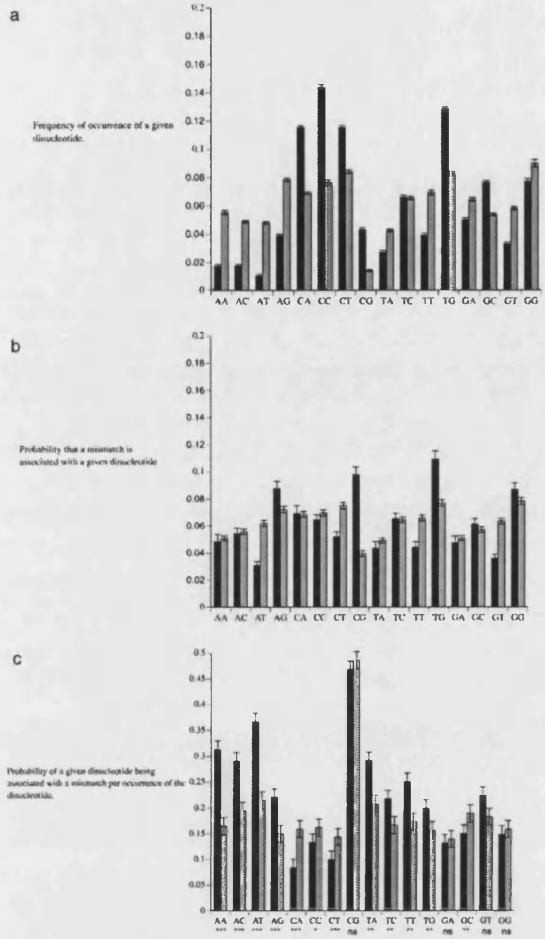


FIG. 2.—At fourfold sites it is the first base of the dinucleotide that occurs at the fourfold site. (*A*) Relative abundance of dinucleotides at fourfold degenerate sites in exons (black bars) and in introns (grey bars). (*B*) Mutability of dinucleotides at fourfold degenerate sites in exons (black bars) and in introns (grey bars). Mutability is defined as the frequency of occurrence of a given dinucleotide in a mismatch. (*C*) Stability of dinucleotides at fourfold degenerate sites in exons (black bars) and in introns (grey bars). Stability is defined as the frequency of occurrence of a given dinucleotide in a mismatch, per incidence of the dinucleotide (i.e., controlling for differences in the relative abundance between fourfold and intronic sites). The significance of differences is indicated by: ns = $P > 0.05$, *$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$.

process of evolution in the two classes are equivalent? Analysis of dinucleotides strongly suggests that we should not. As previously reported in primates (Hellmann et al. 2003; Subramanian and Kumar 2003), CpGs in rodent sequences are more abundant in exons than in introns (fig. 2*A*; CpG content is 7% in exons and 3% in introns). The same is true for all dinucleotides starting with a C. Given the hypermutability of CpG dinucleotides (Bird 1980; McClelland and Ivarie 1982; Cooper and Krawczak 1989; Sved and Bird 1990), our observed excess of TG pairs in exons compared with introns is also expected. The other striking feature is the dearth of the A-starting dinucleotides at fourfold sites in exons.

If all things were equal, we should expect that the probability that a dinucleotide is associated with a mismatch should simply be proportional to the frequency of occurrence of the dinucleotide. However, the C-starting dinucleotides (excluding CpG) are no more likely to be found at a mismatch in exons than in introns (fig. 2*B*). Likewise, mismatches at A-starting dinucleotides (excluding AT) occur at either comparable frequencies in exons and introns or are more abundant in exons, counter to the occurrence of the dinucleotide itself. In other words, the abundance of mismatch-associated dinucleotides per occurrence of the dinucleotide is not the same in exons and introns (fig. 2*C*). Although CpGs are equally likely to be associated with a mismatch in exons and introns, these are the exception. Other C-starting dinucleotides are more "stable" in exons. In contrast, A- and T-starting dinucleotides are more unstable in exons than in introns.

Theoretically, these results could reflect an artifact. Consider a sequence alignment with substitutions randomly located. All things being equal, the number of occurrences of a dinucleotide that is associated with a mismatch could be lower than that of any dinucleotides that happen not to be associated with mismatches. This is simply because association with a mismatch reduces the count of the dinucleotide, as it is present in only one of the two sequences at the site of the mismatch. To exclude this potential bias we reperformed the analysis, but this time we added one to the count of each dinucleotide when it was associated with a mismatch. None of the significant results shown in figure 2*C* are rendered nonsignificant and vice versa.

The apparent instability of A and T at fourfold sites and the opposite apparent stability of C might tempt us to suppose that single nucleotide effects (rather than dinucleotide effects) are the sole interest. However, in exons, the stability of dinucleotides with A or T at the first position is dependent on the nucleotide at the second position (ANOVA on stability of dinucleotides in exons: A-starting dinucleotides, $F = 5.34$, df = 3, $P = 0.001$; T-starting dinucleotides in exons, $F = 4.09$, df = 3, $P = 0.007$; N.B. this result is also found when we adjust for the putative artifact).

Discussion

We have shown that removing putatively constrained sites from introns renders their rate of evolution similar to that of silent sites in the flanking exons. This is consistent with the notion that K_4 and K_i both may be measuring the background mutation rate. However, we have also shown that this similarity in rates of evolution is more likely to be a happy accident, rather than an indication of equivalence in the mode of sequence evolution.

Notably, there are discrepancies in exons and introns in the frequency of occurrence of given dinucleotides associated with mismatches, after controlling for their abundances. Consider, for example, AA and AC. The rate of evolution (mismatch rate) is approximately the same in exons and introns (fig. 2*B*) but only because these

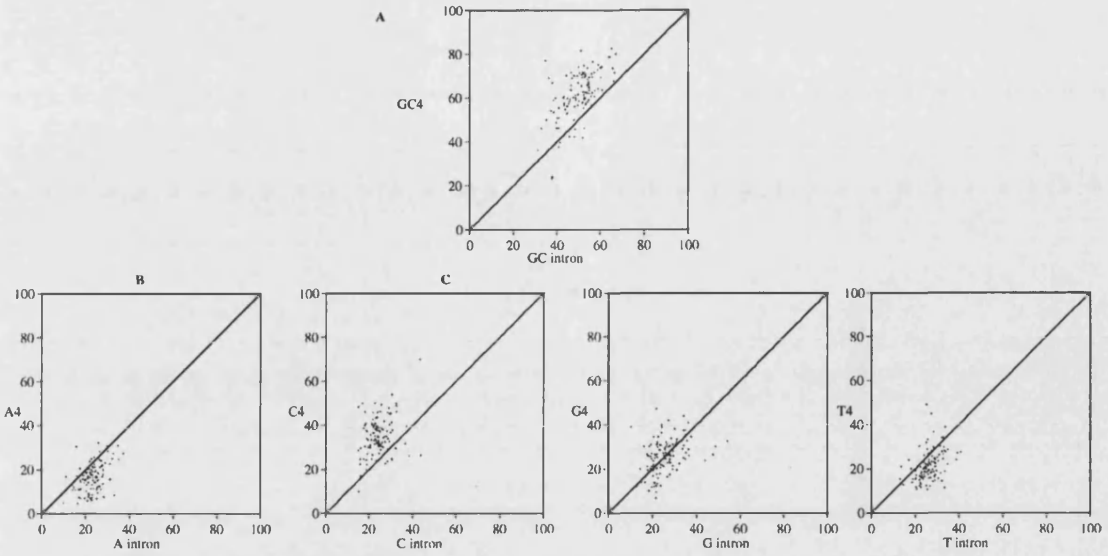


FIG. 3.—Relationship between nucleotide content in introns and at fourfold degenerate sites in flanking exons. (*A*) GC content, (*B*) A content, (*C*) C content, (*D*) G content, and (*E*) T content. The line indicates equality of content.

dinucleotides are both especially rare (fig. 2*A*) and especially unstable in exons (fig. 2*C*). In contrast, the evolution of CpG dinucleotides is more as classically supposed (Cooper and Krawczak 1989; Sved and Bird 1990); the mutability of CpGs appears to be independent of context (intron vs. exon; fig. 2*C*) and the different effects on exons and introns relates solely to different abundances (fig. 2*B*). To account for the latter observation, we need only account for the different CG dinucleotide contents and need not evoke a difference in the mode of evolution of exons and introns. As noted, however, CG is the exception, and only three other dinucleotides (GA, GT, and GG) show the same stability in exons as in introns (fig. 2*C*).

In addition, there is a strand-specific, as well as an exon-specific, enrichment of C nucleotides (fig. 3). Were the effect not strand-specific, we should expect both G and C content in exons to be higher than that in introns. However, C content at fourfold sites is much higher than that in the flanking introns (fig. 3*C*; $P < 0.0001$, one-sample Wilcoxon signed-rank test, $N_g = 136$), while G content is not significantly different between the two (fig. 3*D*; $P = 0.69$, one-sample Wilcoxon signed-rank test, $N_g = 136$). We are not aware of any previous reports of this unexpected difference. These discrepancies cannot be the result of transcription-coupled processes (Green et al. 2003; Majewski 2003), nor to biased gene conversion (Galtier et al. 2001), both of which should affect introns and flanking exons equally. By elimination, we conclude that this is consistent with selectively driven codon usage.

This suggestion is, however, unorthodox. The classical model for selectively driven codon usage bias suggests that its function is to increase the efficiency (rate or accuracy; Duret 2002) of mRNA translation. Evidence for this comes from observations of highly expressed genes having greater bias and that the skews in codon usage reflect iso-acceptor tRNA abundance. Co-adaptation between codon usage, expression rate, and/or tRNA abundance have been described in the worm (Stenico, Lloyd, and Sharp 1994; Duret and Mouchiroud 1999; Duret 2000; Castillo-Davis and Hartl 2002), in the fruitfly (Shields et al. 1988; Moriyama and Powell 1997; Duret and Mouchiroud 1999), and in yeast and bacteria (for reviews, see Ikemura 1985; Sharp et al. 1995), but not in humans (Duret 2002). More generally, in mammals it is usually presumed that the effective population size is too small to allow selection on synonymous codon usage (Sharp et al. 1995) and thus that codon usage bias reflects background isochore GC content (Eyre-Walker 1991; Sharp et al. 1995).

How can we then suggest that there exists selectively driven codon usage, while the abundance of iso-acceptor tRNAs is not skewed (Duret 2002)? Despite this absence of skew, modified patterns of codon usage can affect expression levels in mammals (e.g., Kim, Oh, and Lee 1997). Rather than the product of translational selection, codon usage bias may be the result of selection on mRNA secondary structure (Carlini, Chen, and Stephan 2001), stability, and half-life. Importantly, a recent in vitro study in humans (Duan et al. 2003) has shown that synonymous mutations can be deleterious because of their effect on mRNA secondary structure, reducing stability (see also Gottlieb et al. 1999). Similarly, in bacteria, exchanging synonymous "major" (frequently-used) codons for "minor" ones can result in up to 10-fold reductions in the half-life of mRNA in vivo (Deana and Reiss 1993; Deana, Ehrlich, and Reiss 1996; Deana, Ehrlich, and Reiss 1998). The reduced synonymous substitution rate at the 5′ end of both bacterial (Eyre-Walker and Bulmer 1993) and rodent (Smith and Hurst 1999*b*) coding sequences may also

reflect selection on mRNA secondary structure. The same process may explain, in part, our observation that first introns evolve slowly, even after allowing for the presence of CpG islands (table *B* in the Supplementary Material online).

There are at least two explanations for the effects of silent mutations. On the one hand, the mutations may alter the folding properties, and therefore the stability, of the mRNA. Importantly, several in silico studies report that natural mRNAs (including vertebrate sequences) are more stable than artificial variants that are identical in all regards, other than their synonymous codon usage (Seffens and Digby 1999 [but see Workman and Krogh 1999]; Cohen and Skiena 2003; Katz and Burge 2003). Such arguments fit within the broader context, advocated by Vinogradov (2001*a*, 2001*b*, 2003), that sequence composition reflects selection on physical properties of nucleic acids, such as bendability.

Alternatively, codon usage might alter the ability of mRNA to bind RNA-metabolizing proteins, such as those that bind AU-rich motifs (usually present in 3′ UTRs [Caput et al. 1986; Shaw and Kamen 1986]) and induce rapid mRNA degradation (e.g., Bohjanen et al. 1991). AT-avoidance in exons might therefore minimize the probability of this occurring. The rarest dinucleotide at fourfold degenerate sites in exons is AT (fig. 2*A*), and, given its abundance, it is unusually unstable (fig. 2*C*).

If the preference for C and avoidance of A nucleotides in exons (figs. 2*A* and 3) is the result of selection on mRNA half-life, then we predict that modified versions of mRNAs in mammals, decreasing the abundance of the stable dinucleotides and increasing that of the unstable ones, should, on average, increase mRNA decay rates. For those post-transcriptionally-regulated, this need not be the case (Seffens and Digby 1999). In principle, our prediction could be tested in vitro.

The skew in C but not G content has a bearing on the interpretation of the finding that GC4 is usually greater than GCi (for references, see Duret and Hurst 2001), as also seen in our data set (fig. 3*A*). This is potentially consistent with selection favoring a higher GC content in exons (Hughes and Yeager 1997; Eyre-Walker 1999). This possible interpretation was criticized, as introns also have more GC-poor transposable element (TE) insertions (Duret and Hurst 2001). Vinogradov (2001*b*) counter-argued that TEs cannot account for all of the distortion seen. In addition, that we observe a strand- and exon-dependent enrichment of C, but not of G, is not obviously consistent with the TE model. The model likewise fails to account for the rarity with which C nucleotides feature as a mismatch in exons, given their frequency of occurrence.

Acknowledgments

We thank Martin Lercher, Csaba Pal, and Araxi Urrutia Odabachian for discussion, and we thank Laurent Duret and Jacek Majewski. We are grateful to two anonymous reviewers for insightful comments on an earlier version of the manuscript. J.V.C. and L.D.H. are funded by the United Kingdom Biotechnology and Biological Sciences Research Council.

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Bird, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. **8**:1499–1504.

Bohjanen, P. R., B. Petryniak, C. H. June, C. B. Thompson, and T. Lindsten. 1991. An inducible cytoplasmic factor (AU-B) binds selectively to AUUUA multimers in the 3′ untranslated region of lymphokine mRNA. Mol. Cell Biol. **11**:3288–3295.

Brinster, R. L., J. M. Allen, R. R. Behringer, R. E. Gelinas, and R. D. Palmiter. 1988. Introns increase transcriptional efficiency in transgenic mice. Proc. Natl. Acad. Sci. USA **85**:836–840.

Caput, D., B. Beutler, K. Hartog, R. Thayer, S. Brown-Shimer, and A. Cerami. 1986. Identification of a common nucleotide sequence in the 3′-untranslated region of mRNA molecules specifying inflammatory mediators. Proc. Natl. Acad. Sci. USA **83**:1670–1674.

Carlini, D. B., Y. Chen, and W. Stephan. 2001. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. Genetics **159**:623–633.

Castillo-Davis, C. I., and D. L. Hartl. 2002. Genome evolution and developmental constraint in Caenorhabditis elegans. Mol. Biol. Evol. **19**:728–735.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. **17**:540–552.

Chan, R. Y., C. Boudreau-Lariviere, L. M. Angus, F. A. Mankal, and B. J. Jasmin. 1999. An intronic enhancer containing an N-box motif is required for synapse- and tissue-specific expression of the acetylcholinesterase gene in skeletal muscle fibers. Proc. Natl. Acad. Sci. USA **96**:4627–4632.

Chang, B. H. J., D. Hewett-Emmett, and W.-H. Li. 1996. Male-to-female ratios of mutation-rate in higher primates estimated from intron sequences. Zool. Studies **35**:36–48.

Chang, B. H. J., and W.-H. Li. 1995. Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked *Ube-1* genes and pseudogenes. J. Mol. Evol. **40**:70–77.

Chang, B. H. J., L. C. Shimmin, S. K. Shyue, D. Hewett-Emmett, and W.-H. Li. 1994. Weak male-driven molecular evolution in rodents. Proc. Natl. Acad. Sci. U.S.A. **91**:827–831.

Cohen, B., and S. Skiena. 2003. Natural selection and algorithmic design of mRNA. J. Comput. Biol. **10**:419–432.

Cooper, D. N., and M. Krawczak. 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. **83**:181–188.

Deana, A., R. Ehrlich, and C. Reiss. 1996. Synonymous codon selection controls in vivo turnover and amount of mRNA in Escherichia coli bla and ompA genes. J. Bacteriol. **178**:2718–2720.

———. Silent mutations in the *Escherichia coli* ompA leader peptide region strongly affect transcription and translation in vivo. Nucleic Acids Res. **26**:4778–4782.

Deana, A., and C. Reiss. 1993. Stability of messenger RNA of *Escherichia coli* ompA is affected by the use of synonymous codon. C R Acad. of Sci. III **316**:628–632.

Debry, R. W., and W. F. Marzluff. 1994. Selection on silent sites in the rodent H3 histone gene family. Genetics **138**:191–202.

Dermitzakis, E. T., and A. G. Clark. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. **19**: 1114–1121.

Duan, J., M. S. Wainwright, J. M. Comeron, N. Saitou, A. R. Sanders, J. Gelernter, and P. V. Gejman. 2003. Synonymous

mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum. Mol. Genet. **12**:205–216.

Duret, L. 2000. tRNA gene number and codon usage in the C-elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. **16**:287–289.

———. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12**:640–649.

Duret, L., and L. D. Hurst. 2001. The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Mol. Biol. Evol. **18**:757–762.

Duret, L., and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. Proc. Natl. Acad. Sci. USA **96**:4482–4487.

Duret, L., D. Mouchiroud, and M. Gouy. 1994. HOVERGEN—a database of homologous vertebrate genes. Nucleic Acids Res. **22**:2360–2365.

Eyre-Walker, A. 1991. An analysis of codon usage in mammals: selection or mutation bias? J. Mol. Evol. **33**:442–449.

———. 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics **152**:675–683.

Eyre-Walker, A., and M. Bulmer. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. **21**:4599–4603.

Eyre-Walker, A., and P. D. Keightley. 1999. High genomic deleterious mutation rates in hominids. Nature **397**:344–347.

Fickett, J. W., and A. C. Hatzigeorgiou. 1997. Eukaryotic promoter recognition. Genome Res. **7**:861–878.

Fink, G. R. 1987. Pseudogenes in yeast? Cell **49**:5–6.

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics **159**:907–911.

Gardiner-Garden, M., and M. Frommer. 1987. CpG islands in vertebrate genomes. J. Mol. Biol. **196**:261–282.

Gottlieb, B., D. M. Vasiliou, R. Lumbroso, L. K. Beitel, L. Pinsky, and M. A. Trifiro. 1999. Analysis of exon 1 mutations in the androgen receptor gene. Hum. Mut. **14**:527–539.

Green, P., B. Ewing, W. Miller, P. J. Thomas, NISC Comparative Sequencing Program, and E. D. Green. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. **33**:514–517.

Hare, M. P., and S. R. Palumbi. 2003. High intron sequence conservation across three Mammalian orders suggests functional constraints. Mol. Biol. Evol. **20**:969–978.

Hawkins, J. D. 1988. A survey on intron and exon lengths. Nucleic Acids Res **16**:9893–9908.

Hellmann, I., S. Zollner, W. Enard, I. Ebersberger, B. Nickel, and S. Paabo. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. **13**:831–837.

Huang, W., B. H. J. Chang, X. Gu, D. Hewett-Emmett, and W. H. Li. 1997. Sex differences in mutation rate in higher primates estimated from AMG intron sequences. J. Mol. Evol. **44**:463–465.

Hughes, A. L., and M. Yeager. 1997. Comparative evolutionary rates of introns and exons in murine rodents. J. Mol. Evol. **45**:125–130.

Hural, J. A., M. Kwan, G. Henkel, M. B. Hock, and M. A. Brown. 2000. An intron transcriptional enhancer element regulates IL-4 gene locus accessibility in mast cells. J. Immunol. **165**:3239–3249.

Hurst, L. D., and H. Ellegren. 1998. Sex biases in the mutation rate. Trends Genet. **14**:446–452.

Iida, K., and H. Akashi. 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene **261**:93–105.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.

Jareborg, N., E. Birney, and R. Durbin. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. **9**:815–824.

Jonsson, J. J., M. D. Foresman, N. Wilson, and R. S. McIvor. 1992. Intron requirement for expression of the human purine nucleoside phosphorylase gene. Nucleic Acids Res. **20**:3191–3198.

Kanaya, S., Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. **53**:290–298.

Katai, H., J. D. Stephenson, C. P. Simkevich, J. P. Thompson, and R. Raghow. 1992. An AP-1-like motif in the first intron of human Pro alpha 1(I) collagen gene is a critical determinant of its transcriptional activity. Mol. Cell. Biochem. **118**:119–129.

Katz, L., and C. B. Burge. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. **13**:2042–2051.

Kawada, N., T. Moriyama, A. Ando, T. Koyama, M. Hori, T. Miwa, and E. Imai. 1999. Role of intron 1 in smooth muscle alpha-actin transcriptional regulation in activated mesangial cells in vivo. Kidney International **55**:2338–2348.

Keightley, P. D., and A. Eyre-Walker. 2000. Deleterious mutations and the evolution of sex. Science **290**:331–333.

Kim, C. H., Y. Oh, and T. H. Lee. 1997. Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells. Gene **199**:293–301.

Levy, S., S. Hannenhalli, and C. Workman. 2001. Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics **17**:871–877.

Lothian, C., and U. Lendahl. 1997. An evolutionarily conserved region in the second intron of the human nestin gene directs gene expression to CNS progenitor cells and to early neural crest cells. Eur. J. Neurosci. **9**:452–462.

Majewski, J. 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. Am. J. Hum. Genet. **73**:688–692.

Majewski, J., and J. Ott. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res. **12**:1827–1836.

———. 2003. Amino acid substitutions in the human genome: evolutionary implications of single nucleotide polymorphisms. Gene **305**:167–173.

McClelland, M., and R. Ivarie. 1982. Asymmetrical distribution of CpG in an 'average' mammalian gene. Nucleic Acids Res. **10**:7865–7877.

Moriyama, E. N., and J. R. Powell. 1997. Codon usage bias and tRNA abundance in Drosophila. J. Mol. Evol. **45**:514–523.

Mourier, T., and D. C. Jeffares. 2003. Eukaryotic intron loss. Science **300**:1393–1393.

Oshima, R. G., L. Abrams, and D. Kulesh. 1990. Activation of an intron enhancer within the keratin 18 gene by expression of c-fos and c-jun in undifferentiated F9 embryonal carcinoma cells. Genes Dev. **4**:835–848.

Palmiter, R. D., E. P. Sandgren, M. R. Avarbock, D. D. Allen, and R. L. Brinster. 1991. Heterologous introns can enhance expression of transgenes in mice. Proc. Natl. Acad. Sci. USA **88**:478–482.

Reed, R., and T. Maniatis. 1985. Intron sequences involved in lariat formation during pre-messenger RNA splicing. Cell **41**:95–105.

Rohrer, J., and M. E. Conley. 1998. Transcriptional regulatory elements within the first intron of Bruton's tyrosine kinase. Blood **91**:214–221.

Rossi, P., and B. de Crombrugghe. 1987. Identification of a cell-specific transcriptional enhancer in the first intron of the mouse alpha 2 (type I) collagen gene. Proc. Nat. Acad. Sci. USA **84**:5590–5594.

Sakurai, A., S. Fujimori, H. Kochiwa, S. Kitamura-Abe, T. Washio, R. Saito, P. Carninci, Y. Hayashizaki, and M. Tomita. 2002. On biased distribution of introns in various eukaryotes. Gene **300**:89–95.

Seffens, W., and D. Digby. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res. **27**:1578–1584.

Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi, and J. F. Peden. 1995. DNA-sequence evolution: the sounds of silence. Phil. Trans. R. Soc. Lond. B **349**:241–247.

Shaw, G., and R. Kamen. 1986. A conserved AU sequence from the 3′ untranslated region of GM-CSF mRNA mediates selective mRNA degradation. Cell **46**:659–667.

Shields, D. C., P. M. Sharp, D. G. Higgins, and F. Wright. 1988. Silent sites in Drosophila genes are not neutral—evidence of selection among synonymous codons. Mol. Biol. Evol. **5**:704–716.

Smith, M. W. 1988. Structure of vertebrate genes: a statistical analysis implicating selection. J. Mol. Evol. **27**:45–55.

Smith, N. G. C., and L. D. Hurst. 1998. Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. J. Mol. Evol. **47**:493–500.

———. 1999*a*. The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? Genetics **152**:661–673.

———. 1999*b*. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. Genetics **153**:1395–1402.

Sorek, R., and G. Ast. 2003. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res. **13**:1631–1637.

Stenico, M., A. T. Lloyd, and P. M. Sharp. 1994. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res. **22**:2437–2446.

Subramanian, S., and S. Kumar. 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. **13**:838–844.

Suen, T. C., and P. E. Goss. 2001. Identification of a novel transcriptional repressor element located in the first intron of the human BRCA1 gene. Oncogene **20**:440–450.

Sved, J., and A. Bird. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc. Nat. Acad. Sci. USA **87**:4692–4696.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Mol. Biol. Evol. **10**: 512–526.

Urrutia, A. O., and L. D. Hurst. 2003. The signature of selection mediated by expression on human genes. Genome Res. **13**:2260–2264.

Vinogradov, A. E. 2001*a*. Bendable genes of warm-blooded vertebrates. Mol. Biol. Evol. **18**:2195–2200.

———. 2001*b*. Within-intron correlation with base composition of adjacent exons in different genomes. Gene **276**:143–151.

———. DNA helix: the importance of being GC-rich. Nucleic Acids Res. **31**:1838–1844.

Wasserman, W. W., M. Palumbo, W. Thompson, J. W. Fickett, and C. E. Lawrence. 2000. Human-mouse genome comparisons to locate regulatory sites. Nat. Genet. **26**:225–228.

Waterston, R. H., K. Lindblad-Toh, E. Birney et al. (222 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–562.

Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pruss, I. Reuter, and F. Schacherer. 2000. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. **28**:316–319.

Workman, C., and A. Krogh. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucleic Acids Res. **27**:4816–4822.

Zhuang, Y. A., A. M. Goldstein, and A. M. Weiner. 1989. UACUAAC is the preferred branch site for mammalian mRNA splicing. Proceedings National Academy of Sciences U.S.A. **86**:2752–2756.

Pekka Pamilo, Associate Editor

Accepted December 30, 2003

Online Table A

Differences in putative transcription factor binding site density between first and all non-first introns from the same gene (paired t-tests, $N_g = 116$)

Density	Minimum number of consecutive nucleotides required to define a putative TF binding site			
	2	5	10	15
P-value	0.93	0.23	0.45	0.42
Ratio	1.004	0.964	0.76	0.15
Mean first	0.4540 ± 0.0049	0.1780 ± 0.0043	0.0026 ± 0.0008	0.0006 ± 0.0001
Mean non-first	0.4520 ± 0.0047	0.1850 ± 0.0037	0.0035 ± 0.0008	0.0004 ± 0.0004

Online Table B

Differences in rates of evolution between introns with CpG islands and introns without CpG islands (two tailed t-tests)

Data set / CpG islands?		Minimum number of consecutive CpG dinucleotides required to define a putative CpG island			
		5	10	15	20
	N_i without	458 (94 first)	520 (113 first)	542 (123 first)	549 (127 first)
	N_i with	102 (40 first)	40 (21 first)	18 (11 first)	11 (7 first)
Conservative-K_i	*P*-value	0.05	0.014	0.003	0.008
	K_i without	0.153 ± 0.0021	0.153 ± 0.0019	0.153 ± 0.0019	0.153 ± 0.0018
	K_i with	0.145 ± 0.0038	0.138 ± 0.0056	0.126 ± 0.0076	0.120 ± 0.0096
Liberal-K_i	*P*-value	0.038	0.003	< 0.001	0.002
	K_i without	0.178 ± 0.0028	0.179 ± 0.0029	0.179 ± 0.0025	0.178 ± 0.0025
	K_i with	0.167 ± 0.0049	0.157 ± 0.0065	0.141 ± 0.0089	0.133 ± 0.0100

The number of introns with or without CpG islands that are first introns is given. Note that as the definition of CpG island becomes ever more demanding (A) the number of introns with such control elements goes down, (B) the proportion of those identified as having a CpG island that is also a first intron goes up and (C) the difference in rate of evolution between those with CpG islands and those without becomes all the more evident.

Online Table C

Differences in rates of evolution between first introns without CpG islands and non-first introns also without CpG islands from the same gene (P values from 2 tailed t-test).

Data set		Minimum number of consecutive CpG dinucleotides required to define a putative CpG island			
		5	10	15	20
	N_i first	94	113	123	127
	N_i non-first	364	407	419	422
Conservative-K_i	*P*-value	0.052	0.024	0.024	0.017
	K_i first	0.1460 ± 0.0024	0.145 ± 0.0037	0.146 ± 0.0020	0.145 ± 0.0035
	K_i non-first	0.1553 ± 0.0040	0.155 ± 0.0022	0.155 ± 0.0002	0.155 ± 0.0020
Liberal-K_i	*P*-value	0.019	0.005	0.003	0.001
	K_i first	0.1674 ± 0.0030	0.166 ± 0.0049	0.166 ± 0.0050	0.165 ± 0.0050
	K_i non-first	0.1830 ± 0.0057	0.183 ± 0.0030	0.182 ± 0.0029	0.182 ± 0.0030

Online Table D

Correlation between transcription factor binding site density and intronic substitution rate (analysis done on an intron by intron basis)

Data set		Minimum number of consecutive nucleotides required to define a putative TF binding site			
		2	5	10	15
Conservative-K_i	R^2	0	0	0	0
	P-value	0.59	0.64	0.37	0.95
Liberal-K_i	R^2	0.001	0	0	0
	P-value	0.11	0.62	0.59	0.98
Tight	R^2	0.002	0	0	n/a
	P-value	0.28	0.95	0.70	n/a

Part III. **Understanding biases in synonymous codon usage**

The results of the neutrality test in Part II are consistent with accumulating evidence that synonymous sites are under selection (reviewed in Chapter 7). In Part III, I explore several different mechanisms by which this might occur. In one model, selection affects synonymous codon usage to prevent premature degradation of the mRNA (Chapter 4). The other model posits that there is codon choice to ensure that introns are efficiently spliced-out of the initial transcript (Chapters 5 and 6).

In principle, the C preference I observed at four-fold synonymous sites (Chapter 3) could occur as a consequence of Comeron's (2004) proposed set of preferred codons, of which three-quarters are C-ending. Given, however, that there is still disagreement over whether selection does (Urrutia & Hurst 2003; Comeron 2004; Lavner & Kotlar 2005) or does not (Kanaya et al. 2001; Duret 2002; dos Reis, Savva & Wernisch 2004) maximise the efficiency of protein synthesis in mammals, I decided to test an alternative model.

Both theoretical (Seffens & Digby 1999; Cohen & Skiena 2003) and empirical (Duan et al. 2003; Capon et al. 2004) data have shown that synonymous codon usage can be important for mRNA stability. In Chapter 4, I test the hypothesis (Fitch 1974; Klambt 1975) that selection might act upon synonymous mutations to optimise the thermodynamic stability of mRNA secondary structure. I provide several lines of evidence that supports this idea. Most importantly, the C preference at third (usually four-fold) sites can potentially be explained by selection to favour strong G:C pairs, which increase mRNA stability. This effect may have arisen by virtue of exploiting a tendency for amino acids to use G at the first two sites within codons. Indeed, the stability of wild-type mRNAs relative to artificial transcripts is highest when there is a strong third site skew towards C, and mRNAs are less stable when Gs and Cs are interchanged. Through a novel simulation, I show that, had the synonymous mutations observed in the mouse lineage occurred elsewhere, transcripts would have been less stable. Interestingly, consistent with their proteins being under strong purifying selection, I find that the transcripts of housekeeping genes are also under the greatest pressure to maintain stability.

Is it likely that selection on mRNA stability is the only form of selection on synonymous mutations? Increasing evidence suggests that selection associated with splicing is also important. Two reports have described biases in codon usage that increase as one approaches intron-exon junctions, reflecting selection on codon choice at the pre-mRNA level. Although Willie and Majewski (2004) claimed that the findings

of Eskesen et al. (2004) were compatible with their own, this model of selection is actually two subtly different models that can predict similar effects. The 'cryptic splice site avoidance model' (Eskesen, Eskesen & Ruvinsky 2004) predicts that the gradients in bias are caused by the avoidance of particular codons that might be inappropriately recognised by the intron excision machinery as splice sites. By contrast, the 'enhancer model' hypothesises that specific codons are preferred near intron-exon junctions because they are found in exonic splicing enhancers (ESEs). Curiously, even though the Majewski group had previously described a generalised A+T enrichment at exon ends (Louie, Ott & Majewski 2003; Willie & Majewski 2004), neither they nor Eskesen et al. (2004) attempted to control for this effect. In Chapter 5, I confirm that a generalised A+T enrichment exists, then clarify and test the predictions of the models after controlling for the generalised effect. Overall, I find that there is good support for the enhancer model, but little evidence that codon usage is biased to avoid potential cryptic splice sites.

It is unclear whether the generalised nucleotide bias I observe can easily be explained by either of the models that I described in Chapter 5. The bias may even have nothing to do with selection for splicing efficiency. In Chapter 6, I return to a more direct test for a specific function, namely the well-supported splicing enhancer model. An early study demonstrated that SNP density is lower near intron-exon junctions (Majewski & Ott 2002), which seems to be explained by the presence of ESEs (Fairbrother et al. 2004; Carlini & Genut 2005). Chapter 6 shows that ESEs are under purifying selection, as I find that the synonymous sites in putative ESEs evolve more slowly than the remaining exonic sequence. I also observe that substitutions at four-fold synonymous sites become increasingly less frequent as one approaches the ends of exons, consistent with the trends seen in SNPs. Given the relative abundance of ESEs and the reduced rates of evolution, it appears that the effect of purifying selection on ESEs only leads to around a 10% underestimate the genomic mutation rate estimated from synonymous substitutions.

Chapter 7 features a review of the evidence for selection at synonymous sites and the implications of this finding. In contrast to the discussion from most chapters, which focus on the impact of these results on molecular evolution and underestimating the true point mutation rate, Chapter 7 also describes how selection at synonymous sites will improve our understanding transgenic gene expression and of disease etiology. Lastly, in Chapter 8, I briefly summarise my findings and discuss some of the future perspectives in this field.

References

Capon, F., Allen, M. H., Ameen, M., Burden, A. D., Tillman, D., Barker, J. N. & Trembath, R. C. (2004) A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Hum. Mol. Genet.* **13**: 2361-2368.

Carlini, D. B., & Genut, J. E. (2005) Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Evol.* **In press.**

Cohen, B., & Skiena, S. (2003) Natural selection and algorithmic design of mRNA. *J. Comp. Biol.* **10**: 419-432.

Comeron, J. M. (2004) Selective and mutational patterns associated with gene expression in humans: Influences on synonymous composition and intron presence. *Genetics* **167**: 1293-1304.

dos Reis, M., Savva, R. & Wernisch, L. (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**: 5036-5044.

Duan, J., Wainwright, M. S., Comeron, J. M., Saitou, N., Sanders, A. R., Gelernter, J. & Gejman, P. V. (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* **12**: 205-216.

Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**: 640-649.

Eskesen, S. T., Eskesen, F. N. & Ruvinsky, A. (2004) Natural selection affects frequencies of AG and GT dinucleotides at the 5' and 3' ends of exons. *Genetics* **167**: 543-550.

Fairbrother, W. G., Holste, D., Burge, C. B. & Sharp, P. A. (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* **2**: e268.

Fitch, W. M. (1974) The large extent of putative secondary nucleic acid structure in random nucleotide sequences or amino acid derived messenger-RNA. *J. Mol. Evol.* **3**: 279-291.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. & Ikemura, T. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.* **53**: 290-298.

Klambt, D. (1975) A model for messenger RNA sequences maximizing secondary structure due to code degeneracy. *J. Theor. Biol.* **52**: 57-65.

Lavner, Y., & Kotlar, D. (2005) Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene* **345**: 127-138.

Louie, E., Ott, J. & Majewski, J. (2003) Nucleotide frequency variation across human genes. *Genome Res.* **13**: 2594-2601.

Majewski, J., & Ott, J. (2002) Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12**: 1827-1836.

Seffens, W., & Digby, D. (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.* **27**: 1578-1584.

Urrutia, A. O., & Hurst, L. D. (2003) The signature of selection mediated by expression on human genes. *Genome Res.* **13**: 2260-2264.

Willie, E., & Majewski, J. (2004) Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* **20**: 534-538.

Chapter 4. **Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals**

Jean-Vincent Chamary & Laurence D. Hurst

Genome Biology (2005) **6**: R75

Research

Open Access

Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals

JV Chamary and Laurence D Hurst

Address: Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK.

Correspondence: Laurence D Hurst. E-mail: l.d.hurst@bath.ac.uk

Published: 16 August 2005

Genome Biology 2005, 6:R75 (doi:10.1186/gb-2005-6-9-r75)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/9/R75

Received: 27 April 2005
Revised: 8 June 2005
Accepted: 20 July 2005

Abstract

Background: In mammals, contrary to what is usually assumed, recent evidence suggests that synonymous mutations may not be selectively neutral. This position has proven contentious, not least because of the absence of a viable mechanism. Here we test whether synonymous mutations might be under selection owing to their effects on the thermodynamic stability of mRNA, mediated by changes in secondary structure.

Results: We provide numerous lines of evidence that are all consistent with the above hypothesis. Most notably, by simulating evolution and reallocating the substitutions observed in the mouse lineage, we show that the location of synonymous mutations is non-random with respect to stability. Importantly, the preference for cytosine at 4-fold degenerate sites, diagnostic of selection, can be explained by its effect on mRNA stability. Likewise, by interchanging synonymous codons, we find naturally occurring mRNAs to be more stable than simulant transcripts. Housekeeping genes, whose proteins are under strong purifying selection, are also under the greatest pressure to maintain stability.

Conclusion: Taken together, our results provide evidence that, in mammals, synonymous sites do not evolve neutrally, at least in part owing to selection on mRNA stability. This has implications for the application of synonymous divergence in estimating the mutation rate.

Background

At least in mammals, it is typically assumed that selection does not affect the fate of synonymous (silent) mutations, those nucleotide changes occurring within a gene that affect the coding sequence but not the protein [1,2]. This presumption is in no small part based on the understanding that effective population sizes (N_e) in mammals are small. According to the nearly neutral theory [3], if s is the strength of selection against weakly deleterious mutations, then selection is expected to oppose their fixation when $s > 1/2N_e$ [4]. Consequently, when s is small, species with low N_e are less likely to prevent the fixation of weakly deleterious mutations [5]. Indeed, for species with large effective population sizes, there is little doubt that selection is a strong enough force to determine the fate of synonymous mutations (for example, see [6]). Conversely, in mammals, analyses of codon usage have failed to detect clear signatures of selection (reviewed in [7]).

That synonymous mutations are effectively free of selection is important, not least because, if they really are neutral, their

rate of evolution should be equal to the mutation rate. The rate of synonymous evolution could hence be used to provide a simple and convenient measure of the mutation rate [8,9]. More recently, however, the assumption of neutrality at synonymous sites has been called into question [10-16]. This view has proven contentious, not least because of the absence of a functional role for supposedly silent sites.

Here we examine one hypothesis, that synonymous mutations in mammals are under selection because they affect the thermodynamic stability of mRNA secondary structures [17,18], possibly to prolong cellular half-lives [19,20]. Unlike many non-coding RNAs [21-23], for which a stable secondary structure is selectively favored [24-28], the evolution of a stable structure for mRNA would be constrained by the need to encode a functional protein [17-19,29-31]. Consequently, were selection to operate on mRNA stability, synonymous mutations might be especially important (but see also [32,33]).

The hypothesis is supported by findings that synonymous mutations not only alter mRNA stem-loop structure [34,35], but also affect decay rates, and may lead to disease [35-37]. One possibility is that stem (base-paired) structures protect [38,39] against passive degradation by endoribonucleases [36,40,41]. Similarly, stable structures would be less likely to fall apart and thus expose vulnerable loop (single-stranded) regions to cleavage. Notably, analysis of computationally predicted mRNA stability across a wide taxonomic range revealed that real transcripts are more stable than comparable sequences in which synonymous codons were shuffled while the protein sequence remained unaltered [42,43].

Unfortunately, broad scale empirical analysis of mRNA stability is currently intractable because the structure of sequences much longer than tRNAs cannot be directly observed [20,44]. Consequently, mRNA folding is typically predicted computationally, by one of a variety of methods (see Materials and methods). Importantly, however, no *in silico* method can completely predict how cellular conditions might affect secondary structure [45]. For instance, proteins bound to mature transcripts [46] may have an effect, while chaperones are probably required to guide folding and/or prevent RNAs becoming kinetically trapped in unfavorable conformations [47,48]. Programs that attempt to incorporate the kinetics of the folding process that results from the directionality of transcription [49-51] are still under development [51]. Additionally, although a structure predicted *in silico* might be designated 'correct' because it forms *in vitro*, folding may be somewhat different *in vivo* [48,50].

The premise of this paper is not then to suppose that the prediction method and assumptions are flawless. Rather, we suppose that, if the method is telling us nothing about selection on mRNA stability, there is no reason why multiple independent tests should all point towards the same conclusion. In particular we ask: whether the nucleotides at synonymous sites are non-random with respect to stability; whether the excess of cytosine at synonymous sites in rodents [15] might be accounted for in terms of selection on mRNA stability; whether the location of substitutions in the mouse lineage are non-random with respect to stability; and whether genes under stronger purifying selection also have higher relative stability.

Although the hypothesis predicts that high mRNA stability should be favored, note that we do not expect stability to be extremely high, as ultra-stable structures would impose kinetic barriers that could hinder ribosome translocation [36,52]. While we presume that the transcripts of most genes will be relatively stable, in some cases mRNAs may actually need to be particularly unstable [43]. For example, selection might not act to promote stability because the mRNA is protein-bound and control of expression occurs at the translational level. Alternatively, some genes may only need to be transiently expressed, such as those encoding transcription factors [53,54]. As it is difficult to identify *a priori* which genes these might be, we cannot filter the dataset. This does, however, render our results conservative.

Results

For 70 mouse mRNAs (Additional data file 1), we predict a single optimal putative secondary structure and its thermodynamic stability (ΔG, kcal/mol, the difference in free energy between the folded and unfolded states). Prior studies providing evidence of selection on mRNA structure have employed a randomization protocol that shuffles synonymous codons to generate numerous simulants [42,43,55,56]. Based on the idea that 'interesting' RNAs should be more stable than expected by chance [57], one can then ask whether the stability of a real (wild-type) transcript is, on average, greater than that of its simulants. Seffens and Digby [42], for example, did this for a range of taxa (from bacteria to human). To determine if there is a *prima facie* case to answer, we first performed an analysis similar to that done previously, but specifically restricted to mammalian sequences.

Nucleotide content at synonymous sites is non-random with respect to mRNA stability

If selection acts on synonymous sites, by comparing a real mouse mRNA to simulants differing only at synonymous sites, we should find that, on average, the real transcript is more stable. For each gene we generated 1,000 random mRNAs identical in all regards to the real sequence, but with the bases at 4-fold degenerate (synonymous) sites in the coding sequence (CDS) randomly shuffled between the 4-fold degenerate positions. For each mRNA we determined Z(ΔG), the number of standard deviations the real mRNA is away from the mean stability of the simulants. Z(ΔG) is thus a measure of 'relative stability', the stability of a given mRNA relative to what one would expect by chance alone. Relative

Table 1

Stability of mRNA secondary structures

	Protocol	Mean ΔG	*P*	Mean Z(ΔG)	Mean %pairs
Real (mouse)		-737.98 ± 55.52			60.96 ± 0.28
Modification	Swap G4C4	-734.10 ± 55.08	0.0169		62.11 ± 0.33
Randomization	Sh.4-fold	-725.76 ± 54.71	9e-15	-1.41 ± 0.14	60.77 ± 0.23
	Sh.codon	-728.49 ± 55.01	6e-10	-1.04 ± 0.14	60.61 ± 0.23
	Re-sub.*K*	-733.28 ± 55.15	4e-05	-0.64 ± 0.15	61.06 ± 0.24
	Re-sub.N3	-734.14 ± 55.20	4e-04	-0.51 ± 0.14	61.09 ± 0.24

Means ± SEM are shown, N = 70. *P*-values for modifications were determined by paired t-tests (μ = Real < Modification) on ΔG. *P*-values for randomizations were by one-sample t-tests (expected mean (μ) = 0) on Z(ΔG). %Pairs is the proportion of the coding sequence involved in base-pairing interactions. Artificial sequences generated by the first five protocols encode the same protein as the mouse sequence. A brief description of each protocol follows (see Results for details). 'Sh.4-fold': nucleotides at all 4-fold degenerate sites are shuffled. 'Sh.codon': for each amino acid, the synonymous codons are permuted. 'Re-sub.*K*': synonymous substitutions are reverted back to the rat-mouse common ancestor (rat-mouse common ancestor) state, followed by reallocation of the same number of synonymous point mutations. 'Re-sub.N3': like the previous protocol, except that the nucleotide replacement is also selected at random from the nucleotide distribution at third sites observed in the mouse sequence. 'Swap G4C4': all guanine bases at 4-fold sites are replaced by cytosine, and vice versa.

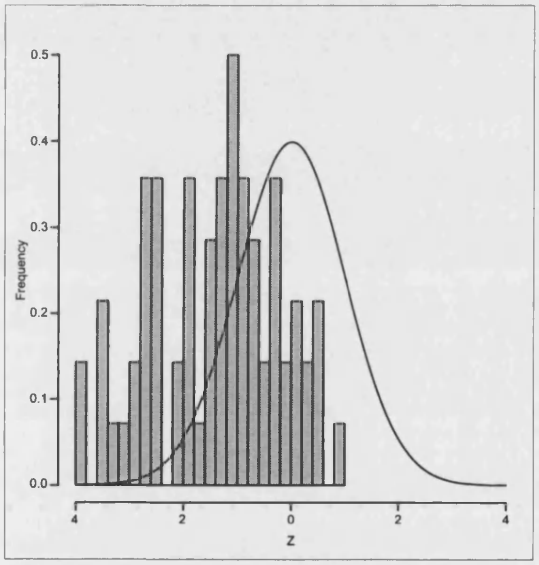


Figure 1
Stability of mRNA secondary structures for 'Sh.4-fold' simulants relative to real transcripts. Histogram of Z-scores for ΔG, the number of standard deviations the real mRNA is away from the mean stability of the simulants, following randomizations shuffling nucleotides at 4-fold degenerate sites (1,000 randomizations per gene, N = 70). The line shows the null normal distribution.

stability can also be considered as a measure of the strength of selection for stability, with a negative Z-score implying higher than expected stability. As Table 1 shows, real mRNAs are, on average, highly significantly more stable than 'Sh.4-fold' simulants (Figure 1; Additional data files 2, 3). Note, however, that on an individual basis, the effect (if any) is weak, with only 26 (37%) of genes having significantly high relative stability at the 5% level (Additional data file 4). Moreover, were we to apply Bonferonni correction for multiple testing on the by-gene *P*-values, no more than four genes would be significant at the 5% level. Inspection of the genes in our dataset (Additional data files 5, 6) did not reveal an obvious pattern that relates relative stability to their function.

In organisms from large effective populations, bias in codon usage is usually attributed to translational selection, favoring efficient (fast and/or accurate) protein synthesis as a consequence of skews in iso-acceptor tRNA abundance (reviewed in [7,58,59]). Whether this occurs in mammals, however, remains a contentious issue. While some have suggested that preferred sets of codons do exist to match the most abundant tRNAs [60], others maintain that codon usage does not reflect tRNAs skews [7,61] and that translational selection does not occur [62]. To be cautious, however, we also employed a protocol ('Sh.codon') that preserves the relative frequency of codons within a given set by shuffling codons within synonymous sets. This protocol gave very similar results to the previous ('Sh.4-fold') randomization (Table 1; Additional data files 2, 3).

Cytosine preference at synonymous sites, diagnostic of selection, can be explained by selection on mRNA stability

While the above results suggest that the identity of the nucleotide at any given synonymous site is non-random, this need not reflect maintenance of mRNA stability. Selection could instead be acting on a thermodynamic property of DNA, such as bendability [63]. As more G:C pairings make helices more bendable and gene-dense regions are GC-rich (for example, see [64]), the putative selection on GC content we observe at

the mRNA level might actually function to provide the transcriptional machinery with easier access to the most gene-dense regions of DNA. To address this issue, we asked about the strand-specific preference for cytosine at 4-fold degenerate sites observed in rodent exons [15].

Cytosine preference is indicated by two related features: a higher C content at 4-fold sites than in flanking introns (not observed for guanine) and an excess of C over G at 4-fold degenerate sites [15]. Correspondingly, we found C4 > G4 in 87% of our mouse genes and a mean skew in GC4 (G - C/G + C) of -0.1506 (P = 1e-11 for expected mean (μ) < 0 by one-sample t-test on GC4 skew). Importantly, the skew towards C is specific to exons and, therefore, cannot be accounted for by effects at the DNA level (for example, mutational biases such as transcription-coupled repair, or selection on transcription). Note also that the sign of the skew is the opposite of that derived from transcription-coupled repair, which yields a G excess [65,66]. Significantly, introducing synonymous changes that increase C|G dinucleotide content (where | is the codon boundary) extends mRNA half-life *in vitro* while increasing A|U enhances degradation [36]. If selection is acting on mRNA stability, then this could be explained by a high C content at third sites increasing the number of potential G:C base-pairs, which are stronger than A:U interactions (triple and double hydrogen bonds, respectively). Consistent with this, we find that genes with the highest relative stability also have a greater excess of C over G (Spearman rank correlation coefficient (ρ) = 0.27, P = 0.0225 for GC4 skew versus $Z(\Delta G^{Sh.4\text{-}fold})$; Additional data file 7).

To further examine the possibility that the C preference is explained by selection on RNA structure, we also asked whether replacing C residues with G decreases stability. We found that real mRNAs are more stable than modified transcripts in which, at 4-fold sites, we swapped all Cs for Gs and vice versa (Table 1; Additional data files 2, 3). 'Swap G4C4' mRNAs, however, possess a higher percentage of base-pairs than real transcripts (62.11 ± 0.33% and 60.96 ± 0.28% in CDS, respectively, P = 0.0003 by paired t-test; 60.84 ± 0.26% and 61.61 ± 0.26% in mRNA, P = 0.0007). That 'Swap G4C4' mRNAs have more base-pairs but lower stability can be explained by the existence of G:U base-pairs within stems, as G:Us are weaker than Watson-Crick interactions (A:U and G:C). An increased G content increases the amount of G:U pairs (real 10.50 ± 0.26% and 'Swap G4C4' 11.64 ± 0.21% in mRNA, P = 3e-07) and thus the proportion of base-paired mRNA, but their stems are less stable (there is no difference in the proportion of A:U pairs: real 36.60 ± 0.70%, 'Swap G4C4' 36.39 ± 0.71%, P = 0.2449). These results further underline the importance of nucleotide content for mRNA rather than DNA stability, not least because the location of bases that can potentially form Watson-Crick base-pairs in DNA is preserved in the modified transcripts.

Biased amino acid content and RNA stability may together drive C preference at third sites

The results above suggest that, given the nucleotide content at non-synonymous sites, C enrichment at synonymous sites is adaptive in regards to mRNA stability. Is there something about non-synonymous sites that causes C in particular to be enriched at synonymous sites? Fitch [17] proposed that, if genetic code degeneracy is exploited to optimize base-pairing in mRNA, third sites within codons (usually synonymous) should be preferentially paired with first and second sites (few and no synonymous sites, respectively). This would also provide a buffer for mRNA structure against non-synonymous substitutions via compensatory changes. Cytosine preference at third sites might, therefore, be driven by selection on amino acid content and mRNA stability [19].

In stems, we expect that, to permit base-pairing, a high G content at first and second sites should be matched by a high C content at third sites (and vice versa), that is, selection on non-synonymous sites would, at least in part, dictate nucleotide content at synonymous sites. At base-paired sites in mRNAs, there is a strong negative correlation in GC skew between first/second sites and third sites (for example, Pearson correlation coefficient (R) = -0.65, P = 1e-09 for GC12 skew versus GC3 skew) that is not observed at unpaired sites (R = 0.70, P = 1e-11; Additional data file 7; note that a positive correlation is expected from isochore structure [67]).

Given the potential inaccuracies of minimum free energy prediction methods (see Materials and methods), we also asked whether the above relationship is robust to the exclusion of sites at which one is less confident that base-pairing occurs (either with a particular site in the optimal structure or with any other site). GC skew is then only calculated for those sites where the probability of pairing is greater than some minimum threshold. We found that the significant negative correlation in GC skew between first/second and third sites is strikingly insensitive to different threshold values (Additional data file 8).

Jia *et al.* [68] recently observed that α-helices and β-sheets of protein secondary structures are preferentially 'coded' by mRNA stems. Using data on the amino acid preferences for protein conformations [69], we found G to be more abundant than C at first and second sites in both α-helices and β-sheets (GC12 skews of 0.001 and 0.0420, respectively). Similarly, there is a bias towards G in these regions within the proteins from our dataset (α-helix GC12 skew of 0.0608 ± 0.0143, P = 8e-05 for μ = 0 by one-sample t-test; β-sheet skew of 0.0879 ± 0.0312, P = 0.0102). The C preference at third sites may, therefore, reflect selection to maintain stable stems in these regions enriched for G at largely non-synonymous sites.

The location of observed synonymous substitutions is non-random with respect to mRNA stability

While randomization protocols that shuffle or swap nucleotides provide insights into how putative selection for mRNA stability and nucleotide content interact, these processes do not occur in nature. The most direct evidence that we can consider is to examine the locations of observed synonymous mutations. Reallocating point mutations is a more realistic form of analysis as it mimics the process of selection following mutation (nucleotide substitutions that are not the result of single point mutations are very rare in mammals, for example, see [70-73]). This minimizes potential biases. For example, randomization protocols that shuffle nucleotides or codons (for example, see [42]) might be problematic [74] as they generate a large number of variants in which there will be a profound effect on dinucleotide relative abundances [75-77]. Simulating the process of evolution, however, only introduces 7 to 8 synonymous changes per 100 sites, hence only about 1 to 2 per 100 nucleotides in the coding sequence. This will have negligible impact on dinucleotide distribution.

Parenthetically, as recent evidence suggests that dinucleotide content in rodent exons is the result of selection [15] and not of biased mutation and/or repair [56,75], the desirability of controlling for dinucleotide distribution is highly questionable. Put differently, if a real mRNA is on average more stable than expected when compared to simulants in which the observed point substitutions have been reallocated, biased dinucleotide distribution is more likely to be a consequence of selection for favorable base-stacking interactions rather than mutational/repair biases.

If certain mutations really were under selection because they diminished mRNA stability, relocating those substitutions actually seen to random locations ('Re-sub.') should lower stability. We used parsimony to determine the substitutions that have arisen in the mouse lineage, inferring the CDS of the rat-mouse common ancestor using hamster as the outgroup to maximize reliability and the number of informative sites (Additional data file 9). We reverted all synonymous changes back to the ancestral state and then simulated mutation by randomly reallocating substitutions at synonymous sites, maintaining the number of observed changes and the encoded protein.

Note that the application of parsimony, while a common practice in the mouse-rat comparison (for example, see [78,79]), can sometimes provide biased ancestral state reconstructions (for example, see [80]). We therefore also reconstructed rat-mouse common ancestor sequences using a maximum likelihood approach. At only 3 of 86,334 reconstructed sites did the parsimony and maximum likelihood methods disagree (excluding sites differing in all three species, see Materials and methods). All three discrepancies occurred in the same gene (*Gadd45a*). Exclusion of this one gene makes no difference to our results (Additional data file 2).

As nucleotide content is influenced by genomic location (isochores; for example, see [67]), the re-introduced nucleotides were selected at random, but in proportion to base composition at third sites in the appropriate mouse gene. This also further minimizes the negligible effect on dinucleotide distributions. From this randomization ('Re-sub.N3') we again find that real mRNAs are, on average, more stable than expected by chance (Table 1; Additional data files 2, 3). Ignoring the effect of isochores and changing the profile of permitted substitutions does not qualitatively alter this result. For example, allowing all mutations to occur with equal likelihood ('Re-sub.*K*') also shows that the locations of observed substitutions have had minimal impact on stability (Table 1; Additional data files 2, 3). Simulants and real transcripts possess a similar amount of base-pairs ($P > 0.15$ by one-sample t-tests on Z(%base-pairs), $\mu = 0$; Table 1).

Signals of selection or methodological artifact?

While the above results indicate that the location at which certain synonymous mutations are observed is in part determined by constraints on mRNA stability, could the above results be artifacts of an inaccurate methodology? We have attempted to minimize such problems by considering those sequences in which *a priori* we expect the method to be more accurate and by considering only those sites that have a high probability of being base-paired. We can, however, consider additional tests. If selection for mRNA stability occurs, we also expect that substitution rates should be related to predicted stem-loop structure and that genes known to be under strong purifying selection should possess mRNAs with high relative stability. We examine these two predictions in turn.

Genes with a high proportion of base-pairs may have fast-evolving stems: evidence for compensatory substitutions?

Testing the first prediction, that evolutionary rates should be linked to mRNA secondary structure, is not straightforward, even if structure prediction were perfect. Although one expects that the majority of compensatory changes will occur to restore substructures, the thermodynamic hypothesis posits that some will also act to restore the overall stability of the molecule. Even if a precise secondary structure were conserved, the difficulty lies in the fact that a given substitution can only be assigned to having occurred within a stem or loop before or after it potentially affects base-pairing, for example, a transversion at a base-paired site in the ancestral mRNA will create a bulge/loop. Consequently, the only substitutions that can be observed within the same (conserved) structure of the descendant sequence are those that arise within loops with little stem-forming potential or within stems in which a compensatory substitution has restored complementary base-pairing. With this caveat in mind, we examined

observed substitutions with respect to the predicted secondary structure in mouse.

We first asked whether substitution rates correlate with the percentage of sequence involved in base-pairing interactions. We found that both the number of synonymous substitutions per synonymous site (Ks) and the non-synonymous substitution rate (Ka) for the whole CDS are higher in genes with more base-pairs (K_a ρ = 0.31, P = 0.0091, N = 70; K_s ρ = 0.31, P = 0.0101, N = 69), although the result for non-synonymous mutations is sensitive to restricting analysis to the subset of small mRNAs (Additional data file 10). These effects seem to be a consequence of substitutional processes within stems. While there is a positive correlation between %base-pairs and rates within putative stems (K_a ρ = 0.31, P = 0.0090, N = 69; K_s ρ = 0.37, P = 0.0020, N = 68), no such relationship exists in loops (K_a ρ = -0.03, P = 0.7941, N = 69; K_s ρ = -0.03, P = 0.8264, N = 69; Additional data file 10).

Note that these latter correlations do not mean that stems evolve faster *per se* (one would predict the opposite), only that they may evolve faster when a lot of the sequence is base-paired. Indeed, consistent with stems being under purifying selection to maintain secondary structure, while non-synonymous rates are the same between codons in putative stems and those in loops (P = 0.6233, N = 69 by paired t-test, stem = 0.0110 ± 0.0018, loop = 0.0095 ± 0.0012), synonymous sites in loops evolve 37% faster than those in stems (P = 0.0045, N = 68, stem = 0.0833 ± 0.0071, loop = 0.0608 ± 0.0034; Additional data file 10).

Why might a high proportion of base-pairing be associated with rapid substitution rates within stems? One possibility is that an abundance of base-pairs ensures that no single mutation can grossly destabilize an mRNA. While one might then predict a negative correlation between %base-pairs and Z(ΔG) (that is, changes to mRNAs with little secondary structure will have a large impact on stability), this may not be observed because when substitutions are randomly reallocated the majority will not fall within stems. Alternatively, the relationship between %base-pairs and substitution rates within stems may indicate a high rate of compensatory changes restoring stem structures. Consider a mutation that arises within a stem that destabilizes the mRNA secondary structure. If selection maintains transcript stability, the substitution will only be tolerated if it is adaptive at the protein level or has such a negligible impact on stability as to be effectively neutral. In the latter case, further changes could accumulate that in combination might significantly alter structure. Under both scenarios, subsequent compensatory mutations restoring stability would thus be under positive selection. The effect of one mutation arising within a stem that has the knock-on effect of increasing substitution rates within stems would be most pronounced in genes with a high proportion of base-pairing. Consequently, compensations would be most favored when there is high pressure to maintain stability. Indeed, we find that in those genes under the strongest selective pressure for high stability, putative stems are fast-evolving (ρ = -0.37, P = 0.0020 for $Z(\Delta G^{Re\text{-}sub.N3})$ versus K_s, N = 68).

Housekeeping genes have high relative stability

To test the second prediction, it is necessary to define *a priori* a set of genes likely to be under stronger purifying selection. Prior evidence indicates that genes expressed in most tissues, housekeeping genes, may be good candidates for two reasons. First, housekeeping proteins evolve slower than tissue-specific ones [73,81-83]. Second, experimental assays of half-life have demonstrated that mRNAs of housekeeping genes degrade relatively slowly [53,54].

Here we identify housekeeping genes by calculating the breadth of expression, the proportion of tissues in which a given gene is expressed. We call a gene 'expressed' in a particular tissue if the average hybridization intensity on microarrays ('average difference' (AD)) for the transcript is greater than 100 or 200 (approximately 2 or 4 copies per cell, respectively, [84]). Housekeeping genes are those expressed in a large proportion of tissues. As described previously (for example, see [73]), we found that protein evolution is slowest in housekeeping genes (%tissues versus K_a: ρ = -0.39, P = 0.0008 for AD > 200; ρ = -0.32, P = 0.0065 for AD > 100).

Significantly, consistent with the prediction, we found that genes subject to strong purifying selection (housekeeping genes) also have the highest relative stability, with the inferred intensity of selection on mRNA stability being correlated with breadth of expression in the expected direction (ρ = -0.25, P = 0.0335 for %tissues versus $Z(\Delta G^{Re\text{-}sub.N3})$ at AD > 200). Using a less conservative cut-off to define a gene as expressed (AD > 100) increases the strength and significance of the correlation (ρ = -0.29, P = 0.0159). The relationship becomes more pronounced after controlling for sequence length (partial ρ = -0.25, P = 0.0179 for AD > 200; partial ρ = -0.30, P = 0.0069 for AD > 100; significance determined by 10,000 randomizations). Expression breadth is not associated with the proportion of the sequence that is base-paired (ρ = -0.01, P > 0.9 for %tissues versus %base-pairs in CDS), nor does the amount of base-pairing predict relative stability (R = -0.14, P = 0.2630 for %base-pairs versus $Z(\Delta G^{Re\text{-}sub.N3})$). As suggested from the 'Swap G4C4' modification protocol, this supports the importance of overall stability over the amount of secondary structure.

Discussion

We have provided numerous lines of evidence that support the hypothesis that selection on synonymous mutations can be mediated by effects on mRNA stability in mammals. Importantly, the signature of selection in rodents, the C preference at 4-fold degenerate sites [15], can potentially be explained by selection on synonymous mutations affecting

mRNA stability. That it should be C in particular (rather than A, G or T), is further explained by skews in nucleotide usage at largely non-synonymous sites: G enrichment at the first and second sites in codons is matched by C enrichment at third sites, so as to ensure, we argue, strong G:C pairs in the mRNA. Moreover, through a randomization that simulates evolution in the mouse lineage, we show that, had the observed substitutions occurred elsewhere within a sequence, they would have had a greater impact on mRNA stability. Additionally, not only do housekeeping genes have unusually low rates of protein evolution, their mRNAs have unusually high relative stability, both features being consistent with stronger selection on this class of genes. Although the structure prediction tool is by no means perfect, it is not obvious how it could be biased in such a way as to cause all our results to point towards the same conclusion.

Synonymous mutations can also be under selection for other functions. Can we be confident that these effects are independent? Recent evidence also suggests that a preference for exonic splicing enhancers (ESEs) affects codon choice [85,86] and that ESEs are under selection [87]. It is likely, however, that the results presented here and selection on ESEs are independent, as ESE hexamers are rich in G compared with C (24% and 14%, respectively, see [86] for dataset), while mRNA stability appears to explain high C content. Moreover, ESEs define relatively little sequence, being short and predominantly located within 20 nucleotides of splice junctions [87].

Experimental predictions for selection on mRNA stability

One might suppose that *in silico* simulations could explain variation in decay rates between genes. Z(ΔG) is not a measure of absolute stability, however, but rather of stability relative to what might have been observed given the underlying parameters of a gene, such as length and coding capacity. Only if all such parameters were equal between genes would one expect relative stability to predict decay rate. However, all else is not equal; for example, we find that Z(ΔG) and nucleotide content covary. Therefore, looking for a correlation between Z(ΔG) and half-life [56] is a weak test because an absence of a relationship would not be strong evidence against the hypothesis unless other variables could be controlled. Indeed, results are ambiguous. Mammalian housekeeping genes have longer half-lives [53,54] and we find that they also have high relative stability. In contrast, Katz and Burge [56] found no correlation between decay rate and local Z(ΔG) in yeast. The interpretation of the yeast result is made even less clear due to uncertainty over when mRNAs should be folded globally. The issue might be easier to resolve once high-quality non-human sequence from primates becomes available, as one could then compare available large-scale surveys of human mRNA decay rates (for example, see [54]) with relative stability. As hominid N_e is around an order of magnitude lower than in murids [88], however, it is also conceivable that selection may not be strong enough to act on mRNA stability.

On the other hand, simulations should predict relative decay rates of mutant versions of a given gene. In at least one case, the dopamine receptor D2 gene, it has been demonstrated that only single nucleotide polymorphisms that induce a conspicuous change in structure predicted *in silico* affect mRNA half-life *in vitro* [35]. A much larger sample set is required to determine whether this is more generally true. We predict that, for those genes with the highest relative stability, the real mRNA should have a longer half-life than the majority of mutants in which one has randomly reallocated synonymous mutations.

Implications for understanding codon usage and mutation rates

That selection maintains mRNA stability contradicts the accepted wisdom that synonymous mutations evolve neutrally [1,2], not only because changes do not alter protein sequence, but also because mammalian effective population sizes (N_e) are thought to be too small to permit selection on mutations of small effect on fitness [6]. Moreover, nucleotide content at silent sites in mammals is best predicted by genomic location (isochores; for example, see [67]). Our observations, however, nonetheless tally with recent evidence that selection acts on synonymous mutations [10-16].

Selection favoring accurate or fast protein synthesis, the classically cited functional role for biased usage of synonymous codons, is not well supported in mammals [7,61,62]. Translational selection predicts that highly expressed genes should exhibit the greatest bias in codon usage [7], but the effect is only weak [13,60,89] and a bias is also observed in lowly expressed genes [89]. On the other hand, selection for mRNA stability need not correlate with expression level (indeed, we find no relationship between Z(ΔG) and mean or peak expression level; $P > 0.1$ in all cases).

When translational selection is known to occur, it can be at odds with selection for mRNA secondary structure (fly, [20]) and stability (yeast, [90]), leading to a trade-off between the two forces [20,90]. Given the difficulties involved in detecting codon usage bias in mammals [7] and our results above, we infer that selection on mRNA stability must be strong relative to translational selection (if the latter occurs at all). This has two repercussions. First, selection for mRNA stability could, in principle, weaken any signal of a preferred set of codons for translational efficiency. Second, in terms of detecting selection at synonymous sites in mammals, asking whether a given amino acid always prefers a certain codon is not necessarily asking the right question. Indeed, it is quite possible that there exist no preferred codon within a gene while at the same time synonymous mutations are under selection. More generally, a complex set of trade-offs between different forms

of selection and mutational biases may render interpretation of patterns of codon usage very difficult.

The evidence for selection on synonymous mutations also has implications for our understanding of both the mutation rate and the mutational load. The substitution rate at synonymous sites in exons is often used as a measure of the mutation rate [8,9]; however, this assumes neutral evolution of synonymous mutations [1,2]. By providing a parsimonious mechanism by which selection could act on synonymous sites, we can ignore the objection that prior evidence is indirect. Nevertheless, it is presently unclear to what degree synonymous mutations are favored or opposed by selection due to their effects on mRNA stability. Without being able to quantify the latter, as well as the net effect of other biases (for example, splice-associated), it will not be possible to directly estimate the extent to which use of the synonymous substitution rate leads to underestimates of the mutation rate and the mutational load.

Conclusion

Recent evidence has suggested that, despite assumptions to the contrary, synonymous mutations in mammalian exons can be under selection. Here we have provided several independent lines of evidence to support the notion that this effect may in part be mediated by selection for mRNA stability. Notably, the preference for cytosine at synonymous sites can be accounted for by such a process. Importantly, the observed substitutions appear to be present at particular sites so as to avoid affecting mRNA stability. Our results have implications for the manner in which codon usage bias should be analyzed to detect selection and for attempts to estimate the mutation rate.

Materials and methods

Orthologous rodent genes

We identified gene families from HOVERGEN (Release 44) [91,92] with complete CDSs for *Mus musculus*, *Rattus norvegicus* and hamster. Orthology was defined as the topology (((mouse, rat), hamster), non-rodent outgroup) within the phylogenetic tree for a given gene, without intervening non-rodent branches between the rodents. Seventy well-described genes matched these criteria and had a <5% size difference between the longest and shortest CDS. Non-redundancy and orthology were supported by syntenic comparisons [93]. Unless otherwise stated, N = 70 for all statistical tests.

Mouse mRNA sequences

Accession numbers from HOVERGEN were used to extract mRNAs from the EnsEMBL genome assembly (Build 30) [94]. When alternative transcripts existed, we used the rat and hamster sequences to identify the desired exons. The untranslated region (UTR) database (Release 15) [95,96] was used for six genes because the UTRs in the EnsEMBL files were unreliably annotated. If present, poly(A) tails were removed as they are coated with binding proteins and so are unlikely to be involved in base-pairing [97].

Coding sequence alignments

Each CDS was extracted using GBPARSE [98] and translated. We aligned amino acid sequences as previously described [15] then reconstructed the three-way nucleotide alignment using AA2NUC (available from L.D.H.).

Reconstruction of rat-mouse common ancestor sequence

Parsimony and maximum likelihood were used to reconstruct ancestral sequence. At 0.3% of sites in the rodent alignment, the rat-mouse ancestral state could not be determined (for example, a different base was present in each species). In these cases, we used the mouse sequence to be conservative for the number of substitutions that have occurred in the mouse lineage. Ancestral states derived from maximum likelihood were determined using codeml in the PAML package [99,100].

RNA secondary structure prediction

There are two main computational approaches to predicting RNA secondary structure. The first is a thermodynamic method, which assumes that a given sequence will fold into the structure with the minimum free energy [101]. The second approach compares multiple orthologous sequences to identify patterns of co-evolution between sites that could be indicative of compensatory mutations [102] to maintain complementary base-pairing within stems [103-108].

In the context of our analysis, the choice is highly constrained and comparative methods may not be applicable to the hypothesis we test. Comparative methods require all input sequences to be of high quality and for the alignment to be accurate. Here we are particularly interested in knowing where substitutions have occurred in a given mammalian lineage and, therefore, need sequence from three species, with mouse-rat-hamster being the obvious choice. Currently, however, rat genomic sequence is not of sufficiently high quality and annotation of UTRs is unreliable. UTRs from hamster are largely unavailable.

Although a moot point under the above circumstances, it may also be undesirable to apply a comparative method in the current context, not least because the logic would be circular: the method requires us to assume that selection is strong enough to maintain secondary structure, while at the same time we are testing for selection. More importantly, based on the evolution of non-coding RNAs, comparative methods are geared towards detecting secondary structure that has been conserved despite sequence divergence [49], that is, well-conserved substructures exist which tend to have specific functions (for example, the anti-codon within a tRNA must always be within a loop). For mRNA, however, a more

realistic model is that selection favors the stability of the mRNA conformation as a whole [17,18]. Highly conserved substructures are not expected *a priori* [109], in part because such conservation may not always be possible, as protein-coding function should outweigh any RNA structure considerations. Essentially, the model assumes that the mRNA will adopt the optimal structure given the available sequence.

Structure and stability were predicted using RNAfold from the Vienna package (Version 1.4) [110,111] under default settings (folding at 37°C, tolerating non-Watson-Crick G:U pairs). Thermodynamic parameters were derived experimentally [112]. RNAfold implements an algorithm that, for a given RNA, finds the conformation with the minimum free energy by maximizing favorable base-pairing interactions [101].

Global versus local mRNA stability

A second methodological issue concerns whether selection might act on stability at the local or global scale. There are two critical issues when choosing which to assess. First, if opposite ends of a molecule are able to pair with one another, RNAs may adopt a conformation closer to a global optimal structure. In eukaryotes, unlike bacteria (where transcription and translation are simultaneous and co-localized), long-range interactions between opposite ends of mRNA molecules can occur [113-116]. This suggests that global [20] rather than local stability is more important to analysis of mammalian sequence.

Second, one must also ask whether the genes contain introns. Generation of a globally stable structure would require the action of spliceosome-associated helicases (for example, [117-119]) to maximize the amount of available sequence. Indeed, it is significant that intronless genes in yeast are less biased for structure than those with introns [56]. All genes in our dataset contain introns, further suggesting global stability to be the more relevant measure. Nonetheless, our assumption of global maximum stability, while an appropriate functional hypothesis, may at best only be a good approximation, as in some cases (for example, short transcripts) there may not be enough time for an mRNA to discover the most optimal structure.

Controling for sequence length

While minimum free energy predictions often agree with laboratory-based methods (for example, stem-loops are avoided at the AUG initiation codon, [120-123]), they are less reliable for long sequences (for example, [112]). The mean length of transcripts in our dataset is 2,101.41 ± 139.84 nucleotides (nt). Consequently, where relevant, we endeavored to control for length effects. In most cases, we carried out the same analyses for mRNAs shorter than 2,000 nt (N = 36, mean mRNA length of 1219.38 ± 77.32 nt), this being the cut-off defining two halves of the dataset. Through Mantell simulations, we found that, when testing for selection on stability, in no instance is the *P*-value for the smaller dataset both not significant and higher than that expected if one were to randomly sample half the dataset, where the full data set analysis suggested significance (Additional data file 3). Consequently we conclude that the results are not obviously biased by the inclusion of long sequence.

Protein function and secondary structure prediction

The attributes of mouse gene products were obtained from the Gene Ontology database (June 2004) [124].

Amino acid sequence was designated as occurring in α-helix, β-sheet (strand) and coil regions using PSIPRED (Version 2.3) [125,126] under default parameters (masking low complexity regions).

Rates of evolution

The number of non-synonymous substitutions per non-synonymous site (K_a) and the synonymous (K_s) distance were estimated with the Li method [127] using the Kimura 2-parameter model. We excluded one fast-evolving gene (K_a = 0.5; K_s = 0.17) in our analyses of evolutionary rates, although inclusion of the outlier gave similar results.

Coding sequence randomization protocols and statistical significance

Simulant mRNAs are identical to their real counterparts in their 5' and 3' untranslated regions and the encoded protein.

On a single-gene basis, the significance of whether its mRNA is more stable than expected by chance is given by:

$$P = \frac{R+1}{N+1}$$

R is the number of artificial mRNAs that are more stable than the real transcript, N is the number (1,000) of randomizations (see Box 1 in [128]).

The Z-score for stability is given by:

$$Z(\Delta G) = \frac{\Delta G^{Real} - \overline{\Delta G}^{Rand}}{\sqrt{\sum_i \frac{(\Delta G_i^{Rand} - \overline{\Delta G}^{Rand})^2}{N-1}}}$$

The Z-scores derived from all randomization protocols are normally distributed.

Expression

Cellular mRNA levels from normalized microarray data on Affymetrix chips were obtained from SymAtlas [129]. We identified the expression profile for each gene by BLASTing mRNA sequences against the probes for the GNF1M chip [130], which has measurements from 61 non-redundant tissue types (the five 'embryo' tissues were ignored). We used the 45-tissue dataset [84] from the U74A chip for six genes

where the suggested BLAST hit from GNF1M were not syntenically feasible. For each tissue we took the mean level across replicate hybridizations. Breadth was set to 0 if AD < 50 in all tissues.

Mantell simulations

To determine whether the incorporation of long genes substantially biased our results, for each modification/randomization protocol, we considered the effect of removing the half of the dataset containing the longest genes. Given that this subset of small mRNAs is by necessity half the size of the full dataset, it is inevitable that *P*-values will be increased. The issue is whether they have increased more than would be expected had we randomly sampled half the dataset. To this end, we randomly extracted 36 genes and re-calculated the significance from t-tests. This was repeated 10,000 times per modification/randomization protocol, yielding the underlying distribution in *P*-values that would be expected were sequence length unimportant. The observed *P*-value (for the shortest genes) was then compared to this expected distribution (see Additional data file 3).

Additional data files

Additional data are available with the online version of this paper. Additional data file 1 contains sequences for all 70 mouse mRNAs in FASTA format. Additional data file 2 is equivalent to Table 1, but excludes the one gene (*Gadd45a*/HBG000516) where the rat-mouse common ancestor sequence differed slightly using the parsimony and maximum likelihood reconstructions. Additional data file 3 is equivalent to Table 1, but only considers mRNAs shorter than 2,000 nucleotides. Additional data file 4 provides the stabilities, relative stabilities and significance values for each modification/randomization on a by-gene basis. Additional data file 5 contains various sequence identifiers (for example, accession numbers) for each mouse gene. Additional data file 6 features gene ontology information, including a description of the function of each mouse gene product. Additional data file 7 contains various correlations for short genes, including GC4 skew versus $Z(\Delta G^{Sh.4\text{-}fold})$, GC12 skew versus GC3 skew (separately for base-paired and unpaired sites) and $Z(\Delta G^{Re\text{-}sub.N3})$ versus K_s at base-paired sites. Additional data file 8 is a table of correlations between GC skew at first/second sites versus skew at third sites, provided for a series of thresholds where the sites analyzed must have a minimum probability of base-pairing. Additional data file 9 is a FASTA file containing three-way alignments of coding sequences from hamster, rat and mouse orthologous genes. Additional data file 10 is a table of correlations for short genes, between the proportion of base-paired sites and non-synonymous or synonymous substitution rates within the coding sequence, base-paired sites and unpaired sites.

Acknowledgements

We thank Csaba Pál for suggesting RNAfold, Fyodor Kondrashov and several anonymous referees for comments. We are also thankful for additional information from the various authors of the programs and databases that were used in this study. J.V.C. is funded by the UK Biotechnology and Biological Sciences Research Council.

References

1. King JL, Jukes TH: **Non-Darwinian evolution.** *Science* 1969, **164**:788-798.
2. Kimura M: **Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution.** *Nature* 1977, **267**:275-276.
3. Ohta T, Gillespie JH: **Development of Neutral and Nearly Neutral Theories.** *Theor Popul Biol* 1996, **49**:128-142.
4. Kreitman M: **The neutral theory is dead. Long live the neutral theory.** *Bioessays* 1996, **18**:678-683.
5. Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF: **DNA sequence evolution: the sounds of silence.** *Philos Trans R Soc Lond B Biol Sci* 1995, **349**:241-247.
6. Shields DC, Sharp PM, Higgins DG, Wright F: **"Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons.** *Mol Biol Evol* 1988, **5**:704-716.
7. Duret L: **Evolution of synonymous codon usage in metazoans.** *Curr Opin Genet Dev* 2002, **12**:640-649.
8. Eyre-Walker A, Keightley PD: **High genomic deleterious mutation rates in hominids.** *Nature* 1999, **397**:344-347.
9. Keightley PD, Eyre-Walker A: **Deleterious mutations and the evolution of sex.** *Science* 2000, **290**:331-333.
10. Iida K, Akashi H: **A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes.** *Gene* 2000, **261**:93-105.
11. Bustamante CD, Nielsen R, Hartl DL: **A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents.** *Mol Biol Evol* 2002, **19**:110-117.
12. Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B, Paabo S: **Selection on human genes as revealed by comparisons to chimpanzee cDNA.** *Genome Res* 2003, **13**:831-837.
13. Urrutia AO, Hurst LD: **The signature of selection mediated by expression on human genes.** *Genome Res* 2003, **13**:2260-2264.
14. Keightley PD, Gaffney DJ: **Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents.** *Proc Natl Acad Sci USA* 2003, **100**:13402-13406.
15. Chamary JV, Hurst LD: **Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively-driven codon usage.** *Mol Biol Evol* 2004, **21**:1014-1023.
16. Lu J, Wu CI: **Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee.** *Proc Natl Acad Sci USA* 2005, **102**:4063-4067.
17. Fitch WM: **The large extent of putative secondary nucleic acid structure in random nucleotide sequences or amino acid derived messenger-RNA.** *J Mol Evol* 1974, **3**:279-291.
18. Klambt D: **A model for messenger RNA sequences maximizing secondary structure due to code degeneracy.** *J Theor Biol* 1975, **52**:57-65.
19. Huynen MA, Konings DA, Hogeweg P: **Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure.** *J Mol Evol* 1992, **34**:280-291.
20. Carlini DB, Chen Y, Stephan W: **The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*.** *Genetics* 2001, **159**:623-633.
21. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, *et al.*: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420**:563-573.
22. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, *et al.*: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome Res* 2004, **14**:331-342.
23. Mattick JS: **RNA regulation: a new genetics?** *Nat Rev Genet* 2004, **5**:316-323.

24. Rivas E, Eddy SR: **Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs.** *Bioinformatics* 2000, **16**:583-605.
25. Wang HC, Hickey DA: **Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes.** *Nucleic Acids Res* 2002, **30**:2501-2507.
26. Bonnet E, Wuyts J, Rouze P, Van De Peer Y: **Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.** *Bioinformatics* 2004, **20**:2911-2917.
27. Meyers LA, Lee JF, Cowperthwaite M, Ellington AD: **The robustness of naturally and artificially selected nucleic acid secondary structures.** *J Mol Evol* 2004, **58**:681-691.
28. Clote P, Ferre F, Kranakis E, Krizanc D: **Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency.** *RNA* 2005, **11**:578-591.
29. White HB 3rd, Laux BE, Dennis D: **Messenger RNA structure: compatibility of hairpin loops with protein sequence.** *Science* 1972, **175**:1264-1266.
30. Ball LA: **Secondary structure and coding potential of the coat protein gene of bacteriophage MS2.** *Nat New Biol* 1973, **242**:44-45.
31. Hasegawa M, Yasunaga T, Miyata T: **Secondary structure of MS2 phage RNA and bias in code word usage.** *Nucleic Acids Res* 1979, **7**:2073-2079.
32. Konecny J, Schoniger M, Hofacker I, Weitze MD, Hofacker GL: **Concurrent neutral evolution of mRNA secondary structures and encoded proteins.** *J Mol Evol* 2000, **50**:238-242.
33. Pedersen JS, Forsberg R, Meyer IM, Hein J: **An evolutionary model for protein-coding regions with conserved RNA structure.** *Mol Biol Evol* 2004, **21**:1913-1922.
34. Shen LX, Basilion JP, Stanton VP Jr: **Single-nucleotide polymorphisms can cause different structural folds of mRNA.** *Proc Natl Acad Sci USA* 1999, **96**:7871-7876.
35. Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, Gejman PV: **Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor.** *Hum Mol Genet* 2003, **12**:205-216.
36. Duan J, Antezana MA: **Mammalian mutation pressure, synonymous codon choice, and mRNA degradation.** *J Mol Evol* 2003, **57**:694-701.
37. Capon F, Allen MH, Ameen M, Burden AD, Tillman D, Barker JN, Trembath RC: **A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups.** *Hum Mol Genet* 2004, **13**:2361-2368.
38. Eichler DC, Eales SJ: **The effect of RNA secondary structure on the action of a nucleolar endoribonuclease.** *J Biol Chem* 1983, **258**:10049-10053.
39. Hambraeus G, Karhumaa K, Rutberg B: **A 5' stem-loop and ribosome binding but not translation are important for the stability of Bacillus subtilis aprE leader mRNA.** *Microbiology* 2002, **148**:1795-1803.
40. Beutler E, Gelbart T, Han JH, Koziol JA, Beutler B: **Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage.** *Proc Natl Acad Sci USA* 1989, **86**:192-196.
41. Qiu L, Moreira A, Kaplan G, Levitz R, Wang JY, Xu C, Drlica K: **Degradation of hammerhead ribozymes by human ribonucleases.** *Mol Gen Genet* 1998, **258**:352-362.
42. Seffens W, Digby D: **mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences.** *Nucleic Acids Res* 1999, **27**:1578-1584.
43. Cohen B, Skiena S: **Natural selection and algorithmic design of mRNA.** *J Comput Biol* 2003, **10**:419-432.
44. Furtig B, Richter C, Wohnert J, Schwalbe H: **NMR spectroscopy of RNA.** *Chembiochem* 2003, **4**:936-962.
45. Gardner PP, Giegerich R: **A comprehensive comparison of comparative RNA structure prediction approaches.** *BMC Bioinformatics* 2004, **5**:140.
46. Dreyfuss G, Kim VN, Kataoka N: **Messenger-RNA-binding proteins and the messages they carry.** *Nat Rev Mol Cell Biol* 2002, **3**:195-205.
47. Herschlag D: **RNA chaperones and the RNA folding problem.** *J Biol Chem* 1995, **270**:20871-20874.
48. Schroeder R, Barta A, Semrad K: **Strategies for RNA folding and assembly.** *Nat Rev Mol Cell Biol* 2004, **5**:908-919.
49. Morgan SR, Higgs PG: **Evidence for kinetic effects in the folding of large RNA molecules.** *J Chem Physics* 1996, **105**:7152-7157.
50. Schroeder R, Grossberger R, Pichler A, Waldsich C: **RNA folding *in vivo*.** *Curr Opin Struct Biol* 2002, **12**:296-300.
51. Meyer IM, Miklos I: **Co-transcriptional folding is encoded within RNA genes.** *BMC Mol Biol* 2004, **5**:10.
52. Kozak M: **Pushing the limits of the scanning mechanism for initiation of translation.** *Gene* 2002, **299**:1-34.
53. Hollams EM, Giles KM, Thomson AM, Leedman PJ: **mRNA stability and the control of gene expression: implications for human disease.** *Neurochem Res* 2002, **27**:957-980.
54. Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell JE Jr: **Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes.** *Genome Res* 2003, **13**:1863-1872.
55. Shpaer EG: **The secondary structure of mRNAs from Escherichia coli: its possible role in increasing the accuracy of translation.** *Nucleic Acids Res* 1985, **13**:275-288.
56. Katz L, Burge CB: **Widespread selection for local RNA secondary structure in coding regions of bacterial genes.** *Genome Res* 2003, **13**:2042-2051.
57. Le SV, Chen JH, Currey KM, Maizel JV Jr: **A program for predicting significant RNA secondary structures.** *Comput Appl Biosci* 1988, **4**:153-159.
58. Ikemura T: **Codon usage and tRNA content in unicellular and multicellular organisms.** *Mol Biol Evol* 1985, **2**:13-34.
59. Akashi H, Eyre-Walker A: **Translational selection and molecular evolution.** *Curr Opin Genet Dev* 1998, **8**:688-693.
60. Comeron JM: **Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence.** *Genetics* 2004, **167**:1293-1304.
61. Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T: **Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis.** *J Mol Evol* 2001, **53**:290-298.
62. dos Reis M, Savva R, Wernisch L: **Solving the riddle of codon usage preferences: a test for translational selection.** *Nucleic Acids Res* 2004, **32**:5036-5044.
63. Vinogradov AE: **Bendable genes of warm-blooded vertebrates.** *Mol Biol Evol* 2001, **18**:2195-2200.
64. Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AH: **The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes.** *Genome Res* 2003, **13**:1998-2004.
65. Green P, Ewing B, Miller W, Thomas PJ, Green ED: **Transcription-associated mutational asymmetry in mammalian evolution.** *Nat Genet* 2003, **33**:514-517.
66. Majewski J: **Dependence of mutational asymmetry on gene-expression levels in the human genome.** *Am J Hum Genet* 2003, **73**:688-692.
67. Eyre-Walker A, Hurst LD: **The evolution of isochores.** *Nat Rev Genet* 2001, **2**:549-555.
68. Jia M, Luo L, Liu C: **Statistical correlation between protein secondary structure and messenger RNA stem-loop structure.** *Biopolymers* 2004, **73**:16-26.
69. Creighton TE: *Proteins: Structure and Molecular Properties* 2nd edition. New York: WH Freeman; 1993.
70. Silva JC, Kondrashov AS: **Patterns in spontaneous mutation revealed by human-baboon sequence comparison.** *Trends Genet* 2002, **18**:544-547.
71. Kondrashov AS: **Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases.** *Hum Mutat* 2003, **21**:12-27.
72. Smith NGC, Webster MT, Ellegren H: **A low rate of simultaneous double-nucleotide mutations in primates.** *Mol Biol Evol* 2003, **20**:47-53.
73. Lercher MJ, Chamary JV, Hurst LD: **Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile.** *Genome Res* 2004, **14**:1002-1013.
74. Workman C, Krogh A: **No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution.** *Nucleic Acids Res* 1999, **27**:4816-4822.
75. Karlin S, Mrazek J: **What drives codon choices in human genes?** *J Mol Biol* 1996, **262**:459-472.
76. Karlin S: **Global dinucleotide signatures and analysis of genomic heterogeneity.** *Curr Opin Microbiol* 1998, **1**:598-610.

77. Gentles AJ, Karlin S: **Genome-scale compositional comparisons in eukaryotes.** *Genome Res* 2001, **11**:540-546.
78. Bazykin GA, Kondrashov FA, Ogurtsov AY, Sunyaev S, Kondrashov AS: **Positive selection at sites of multiple amino acid replacements since the mouse-rat divergence.** *Nature* 2004, **429**:558-562.
79. Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S: **A universal trend of amino acid gain and loss in protein evolution.** *Nature* 2005, **433**:633-638.
80. Eyre-Walker A: **Problems with parsimony in sequences of biased base composition.** *J Mol Evol* 1998, **47**:686-690.
81. Duret L, Mouchiroud D: **Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate.** *Mol Biol Evol* 2000, **17**:68-74.
82. Williams EJB, Hurst LD: **The proteins of linked genes evolve at similar rates.** *Nature* 2000, **407**:900-903.
83. Zhang L, Li WH: **Mammalian housekeeping genes evolve more slowly than tissue-specific genes.** *Mol Biol Evol* 2004, **21**:236-239.
84. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, *et al.*: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci USA* 2002, **99**:4465-4470.
85. Willie E, Majewski J: **Evidence for codon bias selection at the pre-mRNA level in eukaryotes.** *Trends Genet* 2004, **20**:534-538.
86. Chamary JV, Hurst LD: **Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else?** *Trends Genet* 2005, **21**:256-259.
87. Fairbrother WG, Holste D, Burge CB, Sharp PA: **Single nucleotide polymorphism-based validation of exonic splicing enhancers.** *PLoS Biol* 2004, **2**:E268.
88. Keightley PD, Lercher MJ, Eyre-Walker A: **Evidence for widespread degradation of gene control regions in hominid genomes.** *PLoS Biol* 2005, **3**:e42.
89. Lavner Y, Kotlar D: **Codon bias as a factor in regulating expression via translation rate in the human genome.** *Gene* 2005, **345**:127-138.
90. Carlini DB: **Context-dependent codon bias and mRNA longevity in the yeast transcriptome.** *Mol Biol Evol* 2005, **22**:1403-1411.
91. Duret L, Mouchiroud D, Gouy M: **HOVERGEN - a database of homologous vertebrate genes.** *Nucleic Acids Res* 1994, **22**:2360-2365.
92. **HOVERGEN** [http://pbil.univ-lyon1.fr/databases/hovergen.html]
93. **LocusLink** [http://www.ncbi.nlm.nih.gov/LocusLink]
94. **EnsEMBL Mouse Genome Server** [http://www.ensembl.org/Mus_musculus]
95. Pesole G, Liuni S, Grillo G, Licciulli F, Mignone F, Gissi C, Saccone C: **UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002.** *Nucleic Acids Res* 2002, **30**:335-340.
96. **The UTR Database** [http://bighost.area.ba.cnr.it/BIG/UTRHome]
97. Keller RW, Kuhn U, Aragon M, Bornikova L, Wahle E, Bear DG: **The nuclear poly(A) binding protein, PABP2, forms an oligomeric particle covering the length of the poly(A) tail.** *J Mol Biol* 2000, **297**:569-583.
98. **GBPARSE** [http://sunflower.bio.indiana.edu/~wfischer/Perl_Scripts/#gbparse]
99. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13**:555-556.
100. **Phylogenetic Analysis by Maximum Likelihood** [http://abacus.gene.ucl.ac.uk/software/paml.html]
101. Zuker M, Stiegler P: **Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.** *Nucleic Acids Res* 1981, **9**:133-148.
102. Kimura M: **The role of compensatory neutral mutations in molecular evolution.** *J Genet* 1985, **64**:7-19.
103. Dixon MT, Hillis DM: **Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis.** *Mol Biol Evol* 1993, **10**:256-267.
104. Stephan W: **The rate of compensatory evolution.** *Genetics* 1996, **144**:419-426.
105. Higgs PG: **Compensatory neutral mutations and the evolution of RNA.** *Genetica* 1998, **102-103**:91-101.
106. Chen Y, Carlini DB, Baines JF, Parsch J, Braverman JM, Tanda S, Stephan W: **RNA secondary structure and compensatory evolution.** *Genes Genet Sys* 1999, **74**:271-286.
107. Innan H, Stephan W: **Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions.** *Genetics* 2001, **159**:389-399.
108. Savill NJ, Hoyle DC, Higgs PG: **RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods.** *Genetics* 2001, **157**:399-411.
109. Buratti E, Baralle FE: **Influence of RNA secondary structure on the pre-mRNA splicing process.** *Mol Cell Biol* 2004, **24**:10505-10514.
110. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast folding and comparison of RNA secondary structures.** *Monatshefte fur Chimie* 1994, **125**:167-188.
111. **The Vienna RNA Package** [http://www.tbi.univie.ac.at/~ivo/RNA]
112. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.** *J Mol Biol* 1999, **288**:911-940.
113. Konings DA, van Duijn LP, Voorma HO, Hogeweg P: **Minimal energy foldings of eukaryotic mRNAs form a separate leader domain.** *J Theor Biol* 1987, **127**:63-78.
114. Doktycz MJ, Larimer FW, Pastrnak M, Stevens A: **Comparative analyses of the secondary structures of synthetic and intracellular yeast MFA2 mRNAs.** *Proc Natl Acad Sci USA* 1998, **95**:14614-14621.
115. Parsch J, Tanda S, Stephan W: **Site-directed mutations reveal long-range compensatory interactions in the Adh gene of Drosophila melanogaster.** *Proc Natl Acad Sci USA* 1997, **94**:928-933.
116. Parsch J, Stephan W, Tanda S: **Long-range base pairing in Drosophila and human mRNA sequences.** *Mol Biol Evol* 1998, **15**:820-826.
117. Hamm J, Lamond AI: **Spliceosome assembly: the unwinding role of DEAD-box proteins.** *Curr Biol* 1998, **8**:R532-R534.
118. Wang Y, Wagner JD, Guthrie C: **The DEAH-box splicing factor Prp16 unwinds RNA duplexes *in vitro*.** *Curr Biol* 1998, **8**:441-451.
119. Rocak S, Linder P: **DEAD-box proteins: the driving forces behind RNA metabolism.** *Nat Rev Mol Cell Biol* 2004, **5**:232-241.
120. Kozak M: **Influences of mRNA secondary structure on initiation by eukaryotic ribosomes.** *Proc Natl Acad Sci USA* 1986, **83**:2850-2854.
121. de Smit MH, van Duin J: **Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis.** *Proc Natl Acad Sci USA* 1990, **87**:7668-7672.
122. Ganoza MC, Louis BG: **Potential secondary structure at the translational start domain of eukaryotic and prokaryotic mRNAs.** *Biochimie* 1994, **76**:428-439.
123. Rocha EP, Danchin A, Viari A: **Translation in Bacillus subtilis: roles and trends of initiation and termination, insights from a genome analysis.** *Nucleic Acids Res* 1999, **27**:3567-3576.
124. **The Gene Ontology Database** [http://www.geneontology.org]
125. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices.** *J Mol Biol* 1999, **292**:195-202.
126. **The PSIPRED Protein Structure Prediction Server** [http://bioinf.cs.ucl.ac.uk/psipred/psiform.html]
127. Li WH: **Unbiased estimation of the rates of synonymous and nonsynonymous substitution.** *J Mol Evol* 1993, **36**:96-99.
128. Hurst LD, Pal C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order.** *Nat Rev Genet* 2004, **5**:299-310.
129. **SymAtlas** [http://symatlas.gnf.org/SymAtlas]
130. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al.*: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101**:6062-6067.

Additional data file 2

Stability of mRNA secondary structures excluding the Gadd45a gene

	Protocol	Mean ΔG	*P*	Mean Z(ΔG)	Mean %pairs
Real (mouse)		-742.73 ± 56.13			60.99 ± 0.28
Modification	Swap G4C4	-738.79 ± 55.68	0.0170		62.09 ± 0.33
Randomization	Sh.4-fold	-730.24 ± 55.32	2e-15	-1.44 0.14	60.80 ± 0.23
	Sh.codon	-733.06 ± 55.62	3e-10	-1.06 0.14	60.58 ± 0.23
	Re-sub.*K*	-737.89 ± 55.76	2e-05	-0.67 0.15	61.13 ± 0.25
	Re-sub.N3	-738.75 ± 55.81	2e-04	-0.54 0.14	61.16 ± 0.25

Means ±SEM are shown, N=69. *P*-values for modifications are determined by paired t-tests (μ=Real<Mod.) on ΔG. *P*-values for randomisations are by one-sample t-tests (expected mean (μ)=0) on Z(ΔG). %pairs is the proportion of the coding sequence involved in base-pairing interactions. Artificial sequences generated by the first five protocols encode the same protein as the mouse sequence (see Results).

Additional data file 3

Stability of mRNA secondary structures for short genes

	Protocol	Mean ΔG	*P*	Mean Z(ΔG)	Simulated *P*	*P* Mantell	Mean %pairs
Real		-410.90 ±27.36					60.57 ±0.37
Mod.	Swap G4C4	-408.03 ±26.72	0.12		0.1039	0.6938	61.92 ±0.48
Rand.	Sh.4-fold	-402.04 ±26.55	2e-06	-1.22 ±0.21	2e-07	0.9685	60.51 ±0.31
	Sh.codon	-403.42 ±26.67	3e-05	-0.97 ±0.20	5e-05	0.6832	60.25 ±0.30
	Re-sub.*K*	-408.04 ±27.17	0.02	-0.46 ±0.20	0.0126	0.8674	60.63 ±0.35
	Re-sub.N3	-408.50 ±27.18	0.06	-0.37 ±0.19	0.0342	0.8351	60.66 ±0.34

Means ±SEM are shown, N=36. *P*-values for modifications are determined by paired t-tests (μ=Real<Mod.) on ΔG. *P*-values for randomisations are by one-sample t-tests (expected mean (μ)=0) on Z(ΔG). %pairs is the proportion of the coding sequence involved in base-pairing interactions. Artificial sequences generated by the first five protocols encode the same protein as the mouse sequence (see Results). For each Mantell simulation, 10000 datasets of 36 randomly sampled genes were generated (see Materials and methods). Simulated *P* is the mean *P*-value from 10000 t-tests. *P* Mantell is the probability that one would observe a *P*-value less than that obtained from ignoring the long genes due to chance.

Additional data file 7

Miscellaneous correlations for short genes

X	Y	Correlation coefficient	*P*
GC4 skew	$Z(\Delta G^{Sh.4\text{-}fold})$	ρ=0.22	0.20
GC12 skew base-paired	GC3 skew base-paired	*R*=–0.72	8e-07
GC12 skew unpaired	GC3 skew unpaired	*R*=0.67	7e-06
$Z(\Delta G^{Re\text{-}sub.N3})$	K_s base-paired	ρ=–0.40	0.02

N=36, except for $Z(\Delta G^{Re\text{-}sub.N3})$ versus K_s base-paired (N=35). *R*=Pearson correlation coefficient, ρ=Spearman rank correlation coefficient.

Additional data file 8

Relationships between GC12 skew and GC3 skew for a series of minimum base-pairing probabilities

Minimum *P*	*P*(*i*,*j*)			Total *P*(*i*)		
	N	*R*	*P*	N	*R*	*P*
0	70	-0.65	1e-09	70	-0.65	1e-09
0.1	70	-0.46	5e-05	70	-0.60	5e-08
0.2	70	-0.49	1e-05	70	-0.56	4e-07
0.3	70	-0.55	8e-07	70	-0.54	1e-06
0.4	70	-0.57	3e-07	70	-0.52	5e-06
0.5	70	-0.55	7e-07	70	-0.40	0.0005
0.6	69	-0.60	6e-08	69	-0.44	0.0001
0.7	68	-0.52	5e-06	68	-0.38	0.0015
0.8	67	-0.33	0.0061	67	-0.32	0.0073
0.9	64	-0.30	0.0171	66	-0.14	0.2686

GC skew is the bias in cytosine or guanine usage (G-C/(G+C)). *P*(*i*,*j*) is the probability of base-pairing between sites *i* and *j* in the predicted optimal secondary structure. Total *P*(*i*) is the total probability for site *i* being paired with any other site. For each minimum *P*, a given site is only considered when calculating GC skew if its probability of pairing is greater than or equal to the minimum *P*-value.

Additional data file 10

Relationships between the proportion of base-paired sites in mouse coding sequence and rates of evolution for short genes

Site	K_a			K_s		
	Mean ± SEM	ρ	*P*	Mean ± SEM	ρ	*P*
CDS	0.0105 ±0.0019	0.14	0.42	0.0683 ±0.0048	0.41	0.01
Base-paired	0.0097 ±0.0019	0.22	0.21	0.0562 ± 0.0052	0.57	<0.01
Unpaired	0.0127 ±0.0030	–0.08	0.67	0.0895 ± 0.0124	–0.09	0.60

N=35. K_a=non-synonymous substitution rate, K_s= synonymous substitution rate.

ρ=Spearman rank correlation coefficient.

Chapter 5. **Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else?**

Jean-Vincent Chamary & Laurence D. Hurst

Trends in Genetics (2005) **21**: 256-259

8 Ward, I.M. and Chen, J. (2001) Histone H2AX is phosphorylated in an ATR-dependent manner in response to replicational stress. *J. Biol. Chem.* 276, 47759–47762
9 Fernandez-Capetillo, O. *et al.* (2004) H2AX: the histone guardian of the genome. *DNA Repair (Amst.)* 3, 959–967
10 Zou, L. and Elledge, S.J. (2003) Sensing DNA damage through ATRIP recognition of RPA–ssDNA complexes. *Science* 300, 1542–1548
11 Chaudhuri, J. *et al.* (2004) Replication protein A interacts with AID to promote deamination of somatic hypermutation targets. *Nature* 430, 992–998
12 Mol, C.D. *et al.* (1995) Crystal structure and mutational analysis of human uracil-DNA glycosylase: structural basis for specificity and catalysis. *Cell* 80, 869–878
13 Nilsen, H. *et al.* (2000) Uracil-DNA glycosylase (UNG)-deficient mice reveal a primary role of the enzyme during DNA replication. *Mol. Cell* 5, 1059–1065
14 Rada, C. *et al.* (2004) Mismatch recognition and uracil-excision provide complementary paths to both immunoglobulin switching and the second (dA:dT-focussed) phase of somatic mutation. *Mol. Cell* 16, 163–171
15 Bardwell, P.D. *et al.* (2003) The G-U mismatch glycosylase methyl-CpG binding domain 4 is dispensable for somatic hypermutation and class switch recombination. *J. Immunol.* 170, 1620–1624
16 Ehrenstein, M. and Neuberger, M. (1999) Deficiency in Msh2 affects the efficiency and local sequence specificity of immunoglobulin class-switch recombination: parallels with somatic hypermutation. *EMBO J.* 18, 3484–3490
17 Li, Z. *et al.* (2004) Examination of Msh6- and Msh3-deficient mice in class switching reveals overlapping and distinct roles of MutS homologues in antibody diversification. *J. Exp. Med.* 200, 47–59
18 Kolodner, R.D. and Marsischky, G.T. (1999) Eukaryotic DNA mismatch repair. *Curr. Opin. Genet. Dev.* 9, 89–96
19 Sugawara, N. *et al.* (1997) Role of *Saccharomyces cerevisiae* Msh2 and Msh3 repair proteins in double-strand break-induced recombination. *Proc. Natl. Acad. Sci. U. S. A.* 94, 9214–9219

doi:10.1016/j.tig.2005.02.013

Genome Analysis

Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else?

Jean-Vincent Chamary and Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath, UK, BA2 7AY

Two groups recently argued that, in human genes, synonymous sites near intron-exon junctions undergo selection for correct splicing. However, neither study controlled for the possibility of an underlying nucleotide bias at the ends of exons. In this article, we show that generalized A and T enrichment exists, which could be independent of splicing regulation. Evidence for selection between synonymous codons that are associated with splicing enhancers remains after controlling for this bias, whereas support for cryptic splice-site avoidance is diminished.

Introduction

Although synonymous sites are usually thought to evolve neutrally in mammals (e.g. Refs [1,2]), recent evidence suggests otherwise [3–8]. In addition, two groups [9,10] have recently claimed that, near intron-exon junctions, selection for efficient intron excision influences synonymous codon choice. Willie and Majewski [10] reported that, in humans, GAA is more abundant near junctions than GAG, attributing this to the role of GAA as an exonic splicing enhancer (ESE, [11]). This we call the 'enhancer model'. This model fits with evidence that a region of the breast cancer gene, *BRCA1*, with an unusually reduced synonymous substitution rate [12], is a splicing enhancer [13,14] and that, more generally, there is a reduced rate of single nucleotide polymorphisms (SNPs) in ESEs [15]. However, results from Eskesen *et al.* [9] support a different set of hypotheses, which we call the 'cryptic splice-site avoidance model'. They postulated that, because the 3′-ends of introns typically terminate AG, exons should avoid using this dinucleotide at the 5′-end to minimise the chance of deleterious aberrant splice forms.

Disentangling cryptic splice-site avoidance from selection on splicing enhancers

The interpretation of both sets of results is problematic. All of the effects described above have one thing in common: preference for A rather than G near intron-exon junctions. To demonstrate that the suggested interpretations are correct, one must also show that enrichment of A is specific to codons that are associated with splicing, and not a general bias for A. Indeed, a decline in GC content approaching junctions was noted by Majewski's group [10,16], although this might simply reflect GAA preference. Therefore, we first asked whether such a bias exists in codons that are not known to be associated with splicing. Because we found that generalized AT enrichment occurs, the remaining issue is whether we can still find evidence for the forces proposed by the two groups, when controlling for this bias.

Our analysis also attempts to discriminate between the two models. According to the cryptic splice-site avoidance

Corresponding author: Hurst, L.D. (l.d.hurst@bath.ac.uk).
Available online 12 March 2005

model, NAG (N=A|C|G) should be avoided at exonic 5′-ends. By contrast, in the enhancer model, AA preference should be specific to, or most pronounced for, GAA versus GAG, given the role of GAA in ESEs [11]. In principle, the second model could be extended to any codon that is particularly prevalent in ESEs. Similarly, the predictions of the first model should be extended to include conserved sites within exons, which often end with MAG (M=A|C, [17]).

According to the cryptic splice-site avoidance model, because introns start with GT, this dinucleotide should be avoided at exonic 3′-ends. Although Eskesen *et al.* found some evidence for NGT avoidance [9]), part of their analysis compared GT-ending codons with those ending in GC. If GC content decreases approaching junctions, it will be difficult to detect any difference between TGC versus TGT and AGC versus AGT. Comparing CGA with CGT is therefore more informative. The model predicts GT avoidance near the 3′-end, but not near the 5′-end.

To test the predictions, we used a filtered version of the human exon-intron database [18] employed by Eskesen and colleagues. Ignoring first and last exons, the data set consists of 14 407 exons in 1802 genes. For each exon, non-whole codons and the first codon at both ends were ignored. At a given distance from a junction, we measured bias in codon choice by the proportional usage of one codon in a pair of synonymous codons. For each of the 64 possible codons, we calculated their proportional representation in human 5′ and 3′ ESEs from a list of strong candidate hexamers. Pairs of synonymous codons in which both codons are used less than expected by chance (null is approximately ~1.6% usage) or at similar frequencies are useful for examining alternatives to the enhancer model. For more detailed information, see the supplementary material online.

Generalized AT enrichment near junctions might be independent of splicing

To test for T enrichment while controlling for involvement in splicing, we examined the proportional usage of CAT:CAC (both constitute <1.6% of the ESE codon sets) and of GAT:GAC (5′=3.7%:1.6%; 3′=3.2%:2.3%). For both of these codon pairs, T usage increases as one approaches both exon ends (Table 1; supplementary material online). Similarly, a preference for A rather than G near both junctions can be seen by comparing synonyms within CCA:CCG and TTA:TTG pairs (all represent <1.3%), although the decline in A content for TTA:TTG is not significant moving from the 5′-end (Table 1; supplementary material online). We conclude that a generalized AT bias exists near junctions (Figure 1), supporting observations made from a model gene [16]. To be conservative, this should be controlled for to substantiate claims that codons associated with splicing are selectively constrained.

Evidence for GT avoidance is confounded by generalized T enrichment

According to the cryptic splice-site avoidance model, we do not expect any trend in C versus T usage at the 5′–end of exons. However, both AGC:AGT and TGC:TGT (all <0.8%) show T enrichment, further supporting the generalized increased T content near junctions noted previously. The bias for AGT, which is less represented in the ESE codon set than AGC, also suggests that weak enhancer effects do not overpower the T enrichment.

At the 3′-end, interpretation of results is complicated by two confounding factors. First, T enrichment increases GT usage. Second, and conversely, if the Eskesen *et al.* model holds true, minimising the chance of generating cryptic splice sites should drive a decrease in GT usage approaching the 3′-end. The two forces could, in principle, cancel each other out. Indeed, the weak preferences reported in Table 1 for AGT rather than AGC (0.7%:1.6%) and for TGT rather than TGC (0%:0.4%) are not significant. However, the slope of the line for both comparisons is positive, suggesting that AT enrichment might be a stronger force than cryptic GT avoidance.

Although weak AT enrichment is consistent with the idea of opposing forces, can we be certain that GT is ever avoided? The cryptic splice-site avoidance model predicts that CGA will be preferred ahead of CGT at the 3′-end, but not at the 5′-end. General AT enrichment alone predicts no trend at either end. We found that, at the 5′-end, there is no preference for either codon (Table 1). By contrast, T is avoided as one approaches the 3′-end. However, the enhancer model can also predict the same pattern: CGA and CGT are equally used (0.5%) in ESEs with 5′ activity,

Table 1. Proportional useage of one codon in a synonymous pair as a function of distance from intron-exon junctions

		5′-end			3′-end		
Third site	Codon usage	Slope[a]	R^2	P	Slope	R^2	P
A|G	NNA/(NNG+NNA)[b]	−0.0019	0.2755	0.0020	−0.0031	0.2647	0.0025
	AAA/(AAG+AAA)	−0.0013	0.1008	0.0518	−0.0041	0.4521	$4E^{-05}$
	CAA/(CAG+CAA)	−0.0011	0.0929	0.0595	−0.0027	0.3727	0.0003
	GAA/(GAG+GAA)	−0.0044	0.8165	$1E^{-11}$	−0.0038	0.7438	$1E^{-09}$
	CCA/(CCG+CCA)	−0.0025	0.2692	0.0023	−0.0033	0.292	0.0015
	TTA/(TTG+TTA)	−0.0013	0.0392	0.1548	−0.0031	0.137	0.0273
	GGA/(GGG+GGA)	−0.0049	0.5358	$4E^{-06}$	−0.0064	0.6235	$2E^{-07}$
C|T	NNT/(NNC+NNT)[c]	−0.0024	0.4156	0.0004	−0.0020	0.4232	0.0001
	TGC/(TGT+TGC)	+0.0029	0.2536	0.0031	+0.0013	0.045	0.1393
	AGC/(AGT+AGC)	+0.0041	0.5127	$8E^{-06}$	+0.0011	0.0462	0.1365
	GAT/(GAC+GAT)	−0.0022	0.3211	0.0008	−0.0019	0.3447	0.0005
	CAT/(CAC+CAT)	−0.0026	0.2251	0.0054	−0.0023	0.2496	0.0034
A|T	CGA/(CGT+CGA)	−0.0004	0.000	0.6897	−0.0039	0.2099	0.0072

[a]The slope is derived from linear regression weighted by the total number of codons in a synonymous pair at each position.
[b]The mean proportional usage for CCA/(CCG+CCA) and TTA/(TTG+TTA).
[c]The mean proportional usage for CAT/(CAC+CAT) and GAT/(GAC+GAT).

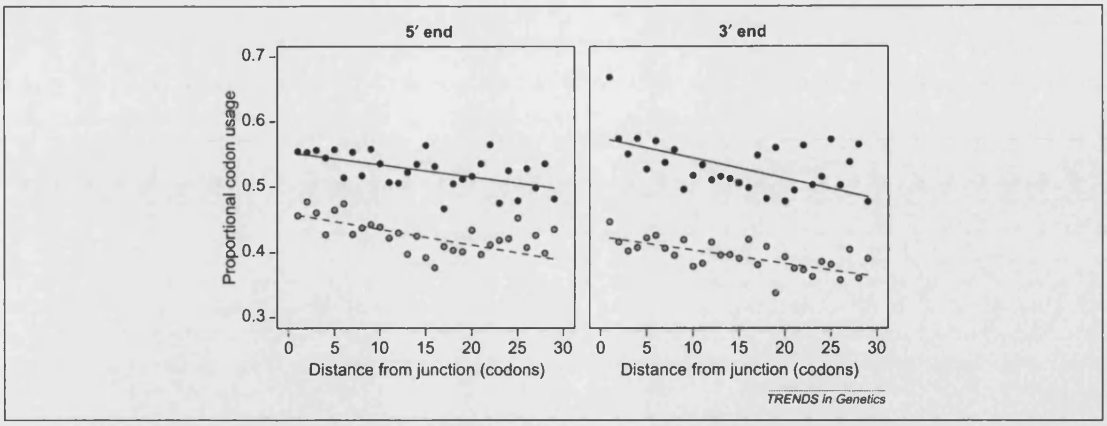


Figure 1. Generalized A and T enrichment near intron-exon junctions. The lines of best fit were derived by regression and weighted by the total number of codons compared at each position. Black circles and the solid lines represent A/(G+A), which is the mean of CCA/(CCG+CCA) and TTA/(TTG+TTA). Open circles and broken lines represent T/(C+T), which is the mean of GAT/(GAC+GAT) and CAT/(CAC+CAT).

but CGT is never used in ESEs at the 3′-end, whereas CGA has a greater than average usage (1.7%).

Evidence for AG avoidance of GAA:GAG but not of MAA:MAG

Eskesen *et al.* found that NAG frequencies increased with distance from the 5′-end of exons [9]. In our extension to their model, AG usage should also be avoided at the 3′-end of exons. By contrast, the enhancer model only predicts biases when synonymous codons are represented to different degrees in candidate ESEs. GAA is a common enhancer [19] and the most prevalent codon in ESEs (5′=11.8%; 3′=11.3%), whereas GAG is represented to a far lesser extent (5′=2.1%; 3′=3.5%). By contrast, AAA and AAG are strongly but equally represented in ESEs (AAA: 5′=7.9%; 3′=7.2%, AAG: 5′=9.2%, 3′=7.5%). CAA is more common in ESEs at the 3′-end of exons (5′=2.4%; 3′=4.7%) than CAG (5′=2.1%; 3′=2.4%). As Table 1 shows, there is a tendency for AA-ending codons to be preferred at both exonic ends. This is seen most profoundly for the GAA:GAG comparison, as predicted by the enhancer model, whereas the effect is not significant for AAA:AAG and CAA:CAG at the 5′-end.

Can these preferences for AA over AG be accounted for by a general A+T bias alone? Using the mean A over G enrichment seen for TTA:TTG and CCA:CCG at the relevant exonic positions as a control, we found that generalized A bias fails to account for enrichment of AA-ending codons at 3′-ends (Table 2), nor can it account for GAA preference at 5′-ends. Preference for AAA over AAG at the 3′-end fits our extension of the cryptic splice-site avoidance model, not least because AAG is more abundant in ESEs than AAA (8.5% and 7.2% respectively). The weak trends seen for AAA:AAG and CAA:CAG at 5′-ends can be explained by generalized A enrichment. This is consistent with the enhancer model but contrary to the cryptic splice-site avoidance model.

Incorporation of first and last exons into the analysis makes no qualitative difference, except that we now observe a weak trend ($P=0.03$) for AAA above AAG at exonic 5′-ends that cannot wholly be accounted for in terms of generalized A enrichment. Whether this reflects a peculiarity of last exons remains to be seen.

What might explain generalized AT enrichment near intron-exon junctions?

The increased AT content near junctions could be unrelated to splicing control. Iida and Akashi [3] show that GC content at the third site in codons for constitutively expressed exons is greater than that of alternatively spliced exons. This might reflect either transcription-coupled mutational biases or selection favouring increased GC content in transcripts that are more frequently translated. Whatever the mechanism, whenever a minor transcript is defined by extension of a constitutive exon, it is the minor form that defines the junction, hence potentially explaining AT bias near junctions.

We cannot, however, be sure that generalized AT bias is independent of splicing, not least because one cannot assume that all exonic enhancers have been identified. Moreover, an AT enrichment at third sites is expected when a synonymous third site is the first position within

Table 2. Proportional usage of synonymous AG-ending codons as a function of distance from intron-exon junctions, controlling for generalized bias for A over G

	5′-end			3′-end		
Codon usage	**Partial *R***	**Partial R^2**	***P*[a]**	**Partial *R***	**Partial R^2**	***P***
AAA/(AAG+AAA)	−0.1763	0.0311	0.1832	−0.5979	0.3575	0.0002
CAA/(CAG+CAA)	−0.0158	0.0003	0.4763	−0.6116	0.374	0.0002
GAA/(GAG+GAA)	−0.8818	0.7776	<0.0001	−0.8671	0.7518	<0.0001
GGA/(GGG+GGA)	−0.7745	0.5999	<0.0001	−0.6787	0.4607	<0.0001

[a] *P*-values for partial correlations determined by 10 000 randomizations.

an ESE because the first nucleotide in known hexamers are commonly A or T (40% A; 28% T). Furthermore, we did not analyse the role of enhancers located within introns, and these should not be present within exons. GGG, for example, is an intronic enhancer [16,20] that is not represented in ESEs, whereas GGA is regularly used (5′=5.8%; 3′=4.5%). Importantly, GGG is relatively rare near exonic boundaries (Table 1), even after correcting for generalized A bias (Table 2). Similarly, the generalized T bias could, in part, result from absence of the (CA)n repeat, an intronic enhancer [21]. However, this would have to occur in the context of at least five alternating ACA and CAC codons.

Concluding remarks

A generalized AT enrichment exists at the ends of exons. This can be independent of splicing enhancers and is almost certainly independent of cryptic splice-site avoidance. Given that there is some degree of uncertainty over the cause, it is conservative to show that any preferences between synonymous codons cannot be accounted for by this bias.

Controlling for AT bias, we find strong evidence that codons that are well-represented in ESEs are preferred. This is seen most profoundly in the GAA:GAG comparison. In our study and that of Willie and Majewski [10], codon biases spread ~30 codons into exons, which also approximately corresponds to the point at which the increasing SNP density in ESEs begins to plateau [15].

The relevance of the cryptic splice-site avoidance model, by contrast, is more questionable. This model, as originally formulated is, at best, consistent with some data but not to the exclusion of the splicing-enhancer model. Both models predict a preference for CAA rather than CAG. Similarly, the evidence for CGA preference at 3′-ends, but not at 5′-ends, is not exclusively supportive of avoiding cryptic GT-splice sites. At the 5′-ends, the lack of strong preference for AAA rather than AAG and CAA rather than CAG suggests that cryptic splice-site avoidance is, at most, a weak effect. However, strong preference for AAA at the 3′-end is contrary to expectations of the enhancer model while supporting our extension of the cryptic splice-site avoidance model. This is the best evidence we have to support any form of cryptic splice-site avoidance.

Acknowledgements

We thank Will Fairbrother for providing the set of candidate ESE clusters, András Kosztolányi for help with creating figures and Anatoly Ruvinsky for helpful discussions. We also thank three anonymous referees for useful suggestions. J.V.C. is funded by the UK Biotechnology and Biological Sciences Research Council.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2005.03.001

References

1 Eyre-Walker, A. and Keightley, P.D. (1999) High genomic deleterious mutation rates in hominids. *Nature* 397, 344–347
2 Keightley, P.D. and Eyre-Walker, A. (2000) Deleterious mutations and the evolution of sex. *Science* 290, 331–333
3 Iida, K. and Akashi, H. (2000) A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene* 261, 93–105
4 Bustamante, C.D. *et al.* (2002) A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Mol. Biol. Evol.* 19, 110–117
5 Hellmann, I. *et al.* (2003) Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* 13, 831–837
6 Urrutia, A.O. and Hurst, L.D. (2003) The signature of selection mediated by expression on human genes. *Genome Res.* 13, 2260–2264
7 Keightley, P.D. and Gaffney, D.J. (2003) Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13402–13406
8 Chamary, J.V. and Hurst, L.D. (2004) Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively-driven codon usage. *Mol. Biol. Evol.* 21, 1014–1023
9 Eskesen, S.T. *et al.* (2004) Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. *Genetics* 167, 543–550
10 Willie, E. and Majewski, J. (2004) Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* 20, 534–538
11 Fairbrother, W.G. *et al.* (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* 297, 1007–1013
12 Hurst, L.D. and Pàl, C. (2001) Evidence for purifying selection acting on silent sites in *BRCA1*. *Trends Genet.* 17, 62–65
13 Liu, H.X. *et al.* (2001) A mechanism for exon skipping caused by nonsense or missense mutations in *BRCA1* and other genes. *Nat. Genet.* 27, 55–58
14 Orban, T.I. and Olah, E. (2001) Purifying selection on silent sites – a constraint from splicing regulation? *Trends Genet.* 17, 252–253
15 Fairbrother, W.G. *et al.* (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* 2, E268
16 Louie, E. *et al.* (2003) Nucleotide frequency variation across human genes. *Genome Res.* 13, 2594–2601
17 Sun, H. and Chasin, L.A. (2000) Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.* 20, 6414–6425
18 Saxonov, S. *et al.* (2000) EID: the exon-intron database-an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.* 28, 185–190
19 Ramchatesingh, J. *et al.* (1995) A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer. *Mol. Cell. Biol.* 15, 4898–4907
20 McCullough, A.J. and Berget, S.M. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell. Biol.* 17, 4562–4571
21 Hui, J. *et al.* (2003) HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nat. Struct. Biol.* 10, 33–37

0168-9525/$ - see front matter
doi:10.1016/j.tig.2005.03.001

Supplementary material

Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice site recognition or something else?

Jean-Vincent Chamary and Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath, UK, BA2 7AY

Corresponding author: Hurst, L.D. (l.d.hurst@bath.ac.uk).

Dataset of complete coding sequences

We started with the same version of the human exon-intron database employed by Eskesen *et al.* [1], which contains 47 908 exons from 7150 entries. However, many sequences are incomplete, not starting with ATG and/or not finishing with a stop codon. These were eliminated. Sequences with non-terminal in-frame stop codons were also excluded. Moreover, many genes are duplicates of some variety, either different but similar genes, different versions of the same gene or, in some cases, identical copies of the same gene. We therefore performed an all-against-all BLAST (using E = 0.001), eliminating all but one of each duplicate cluster. We arbitrarily retained the longest entry or, when several long coding sequence were present, selected one at random. Of the remaining coding sequences, we cross-referenced the protein identifier with NCBI (http://www.ncbi.nlm.nih.gov), which revealed five non-human genes. Before ignoring first and last exons, this left 18 414 exons from 2033 genes. Our final dataset consisted of 14 407 exons in 1802 genes.

Proportional usage of synonymous codon pairs as a function of distance from junctions

All exons were trimmed so that the first and last codons were whole (i.e. removing 0–2 nucleotides from both ends). Each exon was divided in two, the first half being considered the 5'-end, the second the 3'-end. Under this protocol no given codon can be counted more than once. Running towards the interior of an exon, the distance from the intron-exon junction is the number of whole codons between a given codon and the junction pertinent to the half-exon. The first whole codon at each end was hence the 0^{th} codon.

For each synonymous codon pair of interest, we count the number of codons at each position relative to the intron-exon junction. We estimate biases by considering the proportional use of the first codon within the pair, which is given by codon1/(codon1 + codon2). The 0^{th} codon was excluded due to known preferences at junctions (e.g. because AAG is often the last codon at the 3' end of an exon [2], AAA will be comparatively rare).

Proportional representation of codons in candidate exonic splicing enhancer hexamers

We counted the number of occurrences of each of the 64 possible codons in a list of strong candidate hexamers, separated into clusters with high sequence similarity (kindly provided by Will Fairbrother, [3]) that define human ESEs. From these clusters we determined whether a given hexamer enhanced splicing at the 5' and/or 3'-ends of exons. We then split each hexamer in the non-redundant lists into four codons (starting at positions 1, 2 and 3), which yielded 95 hexamers for the 5'-end (a 380-codon set) and 177 hexamers with 3' activity (708-codon set). A list of 238 hexamers that is not split by 5' or 3' activity can be obtained from the RESCUE-ESE web server (http://genes.mit.edu/burgelab/rescue-ese).

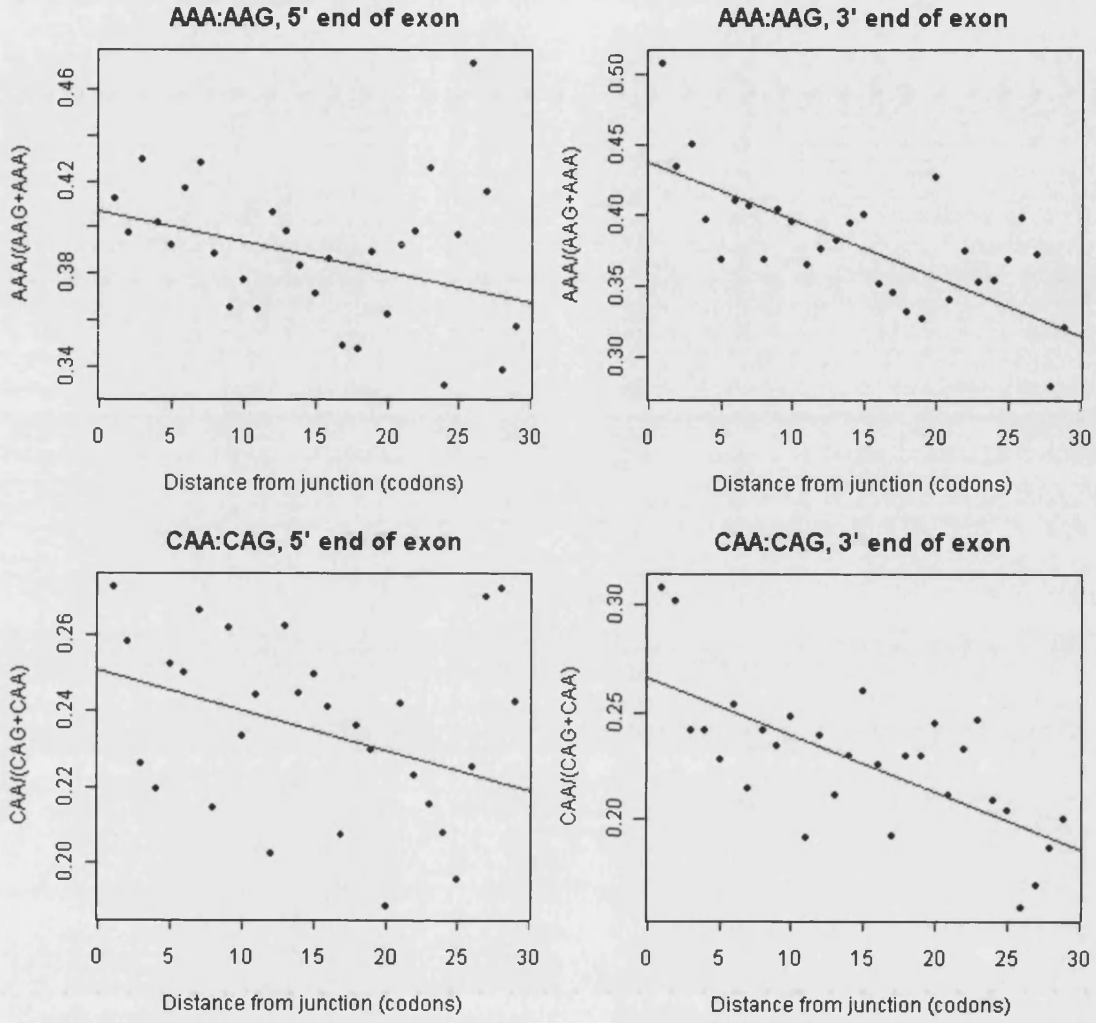
References

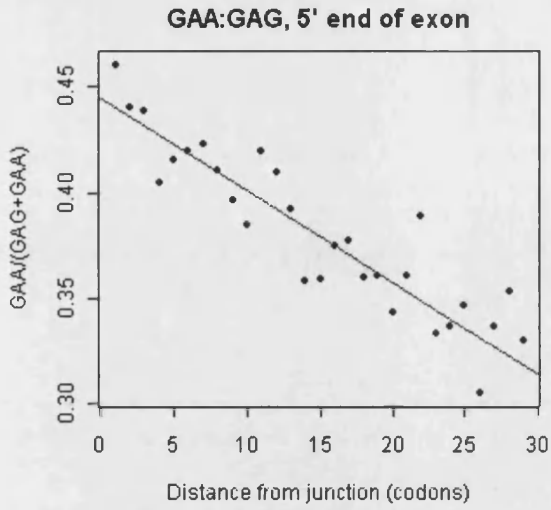
1 Eskesen, S.T. *at al.* (2004) Natural selection affects frequencies of AG and GT dinucleotides at the 5' and 3' ends of exons. *Genetics* 167, 543–550

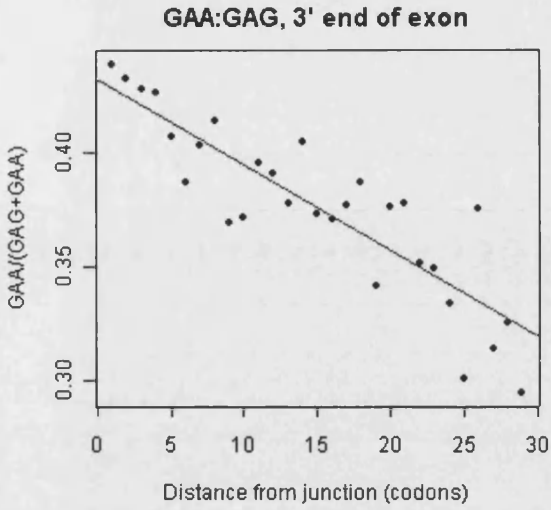
2 Sun, H. and Chasin, L.A. (2000) Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.* 20, 6414–6425

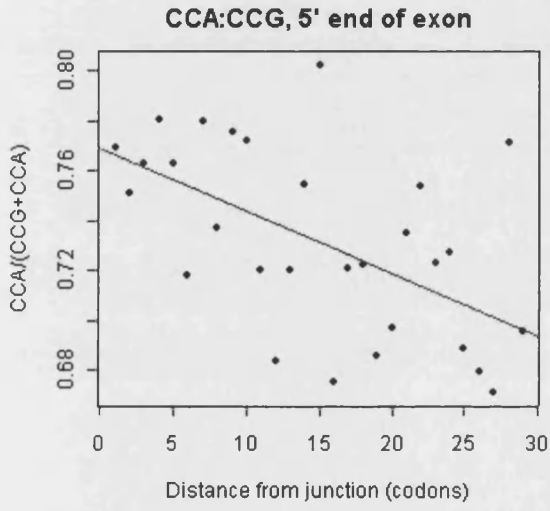
3 Fairbrother, W.G. *et al.* (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* 297, 1007–1013

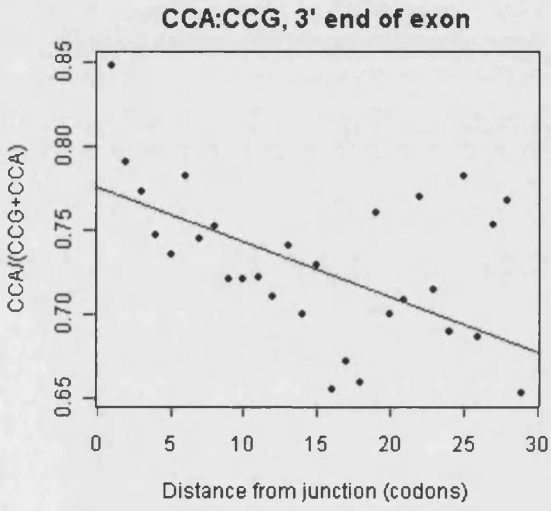
Figure 1. The proportional usage of one codon within a synonymous pair as a function of distance from intron-exon junctions. For each of the following synonymous codon pairs, the plot for the distance from the 5'-end (in the 5' to 3' direction) of the exon is plotted on the left hand side, whereas distance from the 3'-end of the exon (in the 3' to 5' direction) is plotted on the right hand side. The plotted lines are the best fit regression lines weighted by the number of codons in the sample, for each given codon pair, at the given distance. For a statistical resume of results, see Table 1 in the main text.

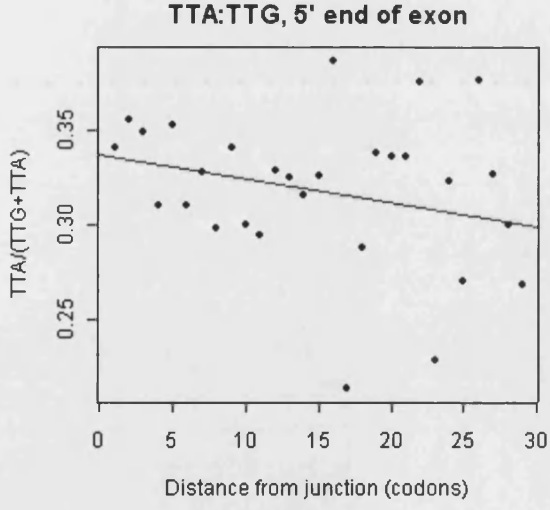


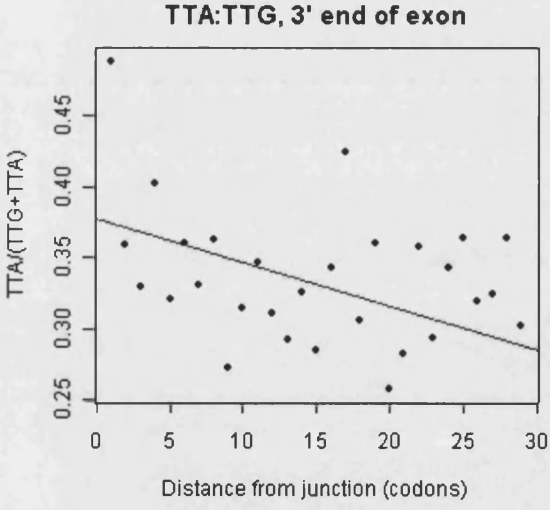

GAA:GAG, 5' end of exon
GAA/(GAG+GAA)
0.30
0.35
0.40
0.45
0
5
10
15
20
25
30
Distance from junction (codons)

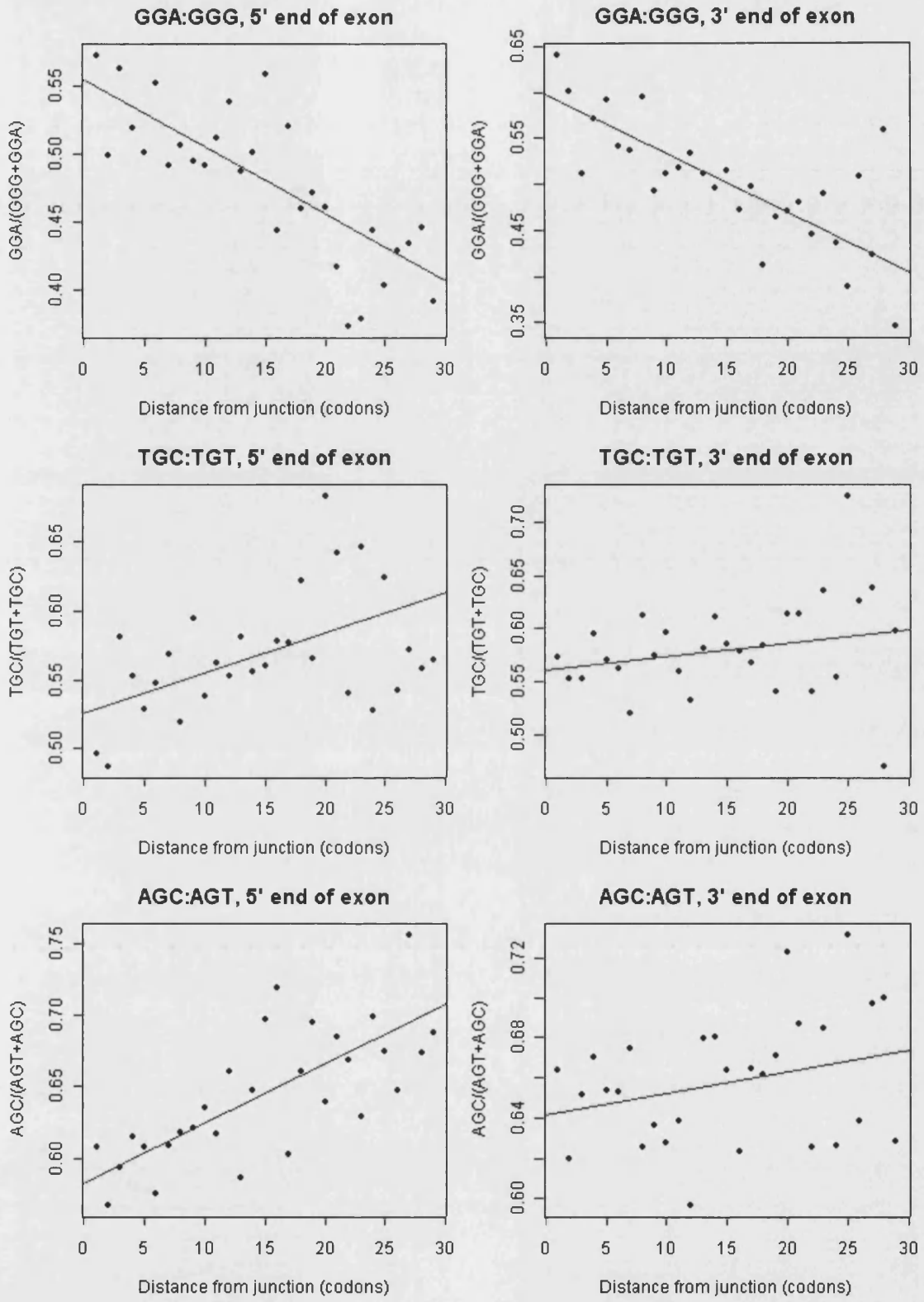

GAA:GAG, 3' end of exon
GAA/(GAG+GAA)
0.30
0.35
0.40
0
5
10
15
20
25
30
Distance from junction (codons)

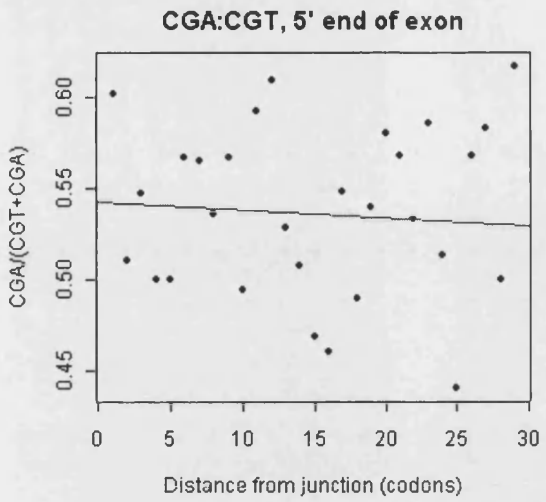

CCA:CCG, 5' end of exon
CCA/(CCG+CCA)
0.68
0.72
0.76
0.80
0
5
10
15
20
25
30
Distance from junction (codons)

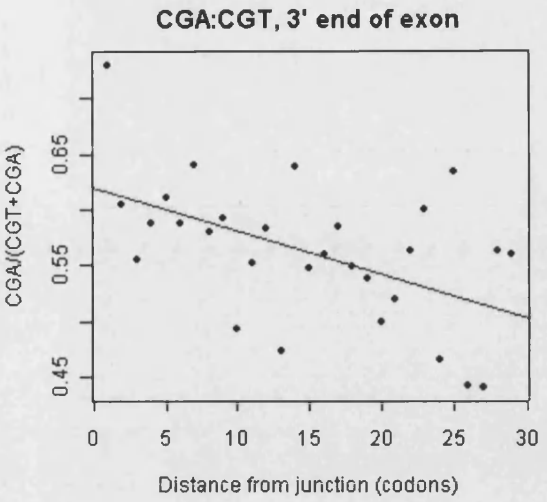

CCA:CCG, 3' end of exon
CCA/(CCG+CCA)
0.65
0.70
0.75
0.80
0.85
0
5
10
15
20
25
30
Distance from junction (codons)

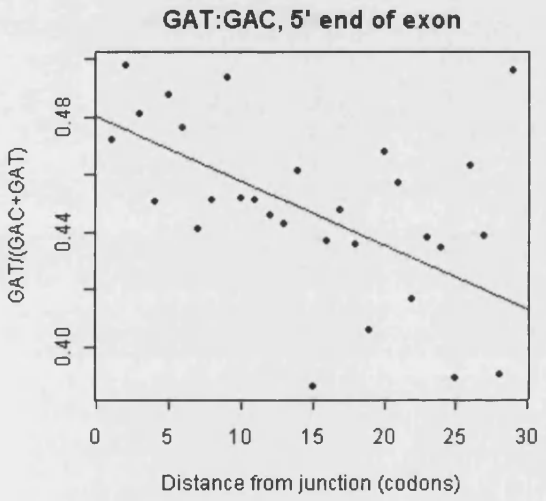

TTA:TTG, 5' end of exon
TTA/(TTG+TTA)
0.25
0.30
0.35
0
5
10
15
20
25
30
Distance from junction (codons)

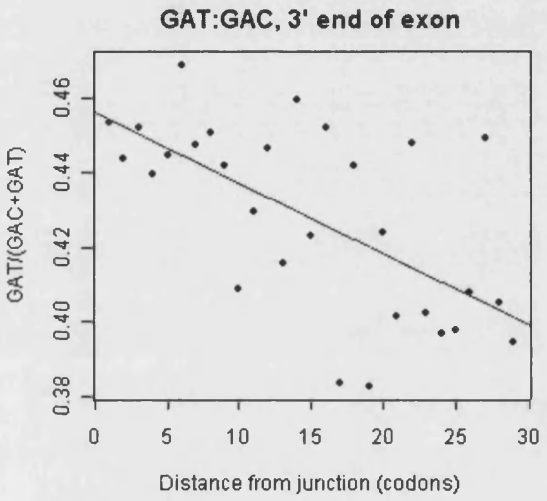

TTA:TTG, 3' end of exon
TTA/(TTG+TTA)
0.25
0.30
0.35
0.40
0.45
0
5
10
15
20
25
30
Distance from junction (codons)

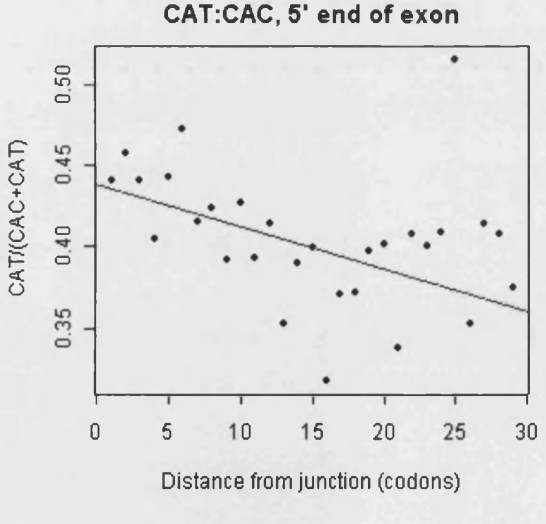

GGA:GGG, 5' end of exon
GGA/(GGG+GGA)
0.40
0.45
0.50
0.55
0 5 10 15 20 25 30
Distance from junction (codons)
GGA:GGG, 3' end of exon
GGA/(GGG+GGA)
0.35
0.45
0.55
0.65
0 5 10 15 20 25 30
Distance from junction (codons)
TGC:TGT, 5' end of exon
TGC/(TGT+TGC)
0.50
0.55
0.60
0.65
0 5 10 15 20 25 30
Distance from junction (codons)
TGC:TGT, 3' end of exon
TGC/(TGT+TGC)
0.50
0.55
0.60
0.65
0.70
0 5 10 15 20 25 30
Distance from junction (codons)
AGC:AGT, 5' end of exon
AGC/(AGT+AGC)
0.60
0.65
0.70
0.75
0 5 10 15 20 25 30
Distance from junction (codons)
AGC:AGT, 3' end of exon
AGC/(AGT+AGC)
0.60
0.64
0.68
0.72
0 5 10 15 20 25 30
Distance from junction (codons)

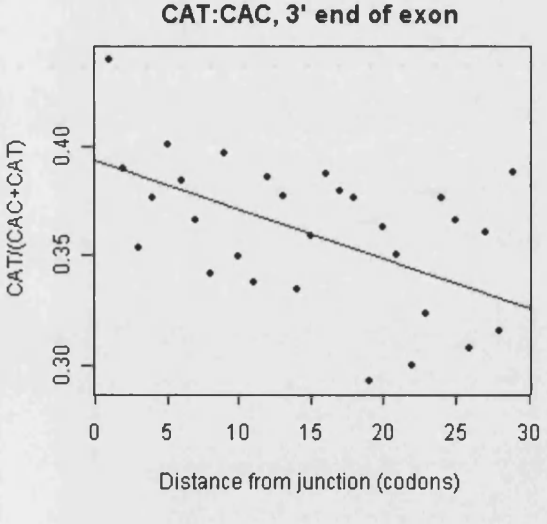

CGA:CGT, 5' end of exon
CGA/(CGT+CGA)
0.45
0.50
0.55
0.60
0
5
10
15
20
25
30
Distance from junction (codons)


CGA:CGT, 3' end of exon
CGA/(CGT+CGA)
0.45
0.55
0.65
0
5
10
15
20
25
30
Distance from junction (codons)


GAT:GAC, 5' end of exon
GAT/(GAC+GAT)
0.40
0.44
0.48
0
5
10
15
20
25
30
Distance from junction (codons)


GAT:GAC, 3' end of exon
GAT/(GAC+GAT)
0.38
0.40
0.42
0.44
0.46
0
5
10
15
20
25
30
Distance from junction (codons)


CAT:CAC, 5' end of exon
CAT/(CAC+CAT)
0.35
0.40
0.45
0.50
0
5
10
15
20
25
30
Distance from junction (codons)


CAT:CAC, 3' end of exon
CAT/(CAC+CAT)
0.30
0.35
0.40
0
5
10
15
20
25
30
Distance from junction (codons)

Chapter 6. **Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers**

Joanna L. Parmley, Jean-Vincent Chamary & Laurence D. Hurst

Molecular Biology and Evolution (accepted pending minor revision)

Silent sites in mammals have classically been considered to be free from selective pressures. Under this assumption, the synonymous substitution rate is used as a proxy for the mutation rate. Accumulating evidence, however, suggests that the assumption is not valid, although the mechanism(s) by which selection acts remains unclear. Recent work has revealed the potential importance of exonic splicing enhancers (ESEs) in synonymous evolution. ESEs are predominantly located near intron-exon junctions, which may explain the reduced SNP density in these regions. Here we show that synonymous sites in putative ESEs evolve at a significantly lower rate than the remaining exonic sequence. Differential mutabilities of ESEs do not appear to explain this difference. We observe that substitution frequency at four-fold degenerate sites decreases as one approaches the ends of exons, consistent with the existing SNP data. This trend is at least in part explained by lower substitution rates in ESEs coupled with their relative abundance near junctions. Given the relative abundance of ESEs and the reduced rates of evolution, we estimate that use of synonymous divergence underestimates the mutation rate not more than around 8%, assuming no other causes of selection on synonymous sites. Selection on exonic splicing enhancers, in addition, appears to affect non-synonymous evolution and amino acid usage near intron-exon junctions is also non-random.

Introduction

At least in mammals, synonymous (silent) sites have long been assumed to be free from the pressures of natural selection (Eyre-Walker 1991; Sharp et al. 1995). If synonymous mutations are neutral (Kimura 1983) then the rate of synonymous divergence can be employed to measure the point mutation rate (Eyre-Walker and Keightley 1999; Keightley and Eyre-Walker 2000). Recently, however, there has been mounting evidence against this line of thought (Iida and Akashi 2000; Bustamante, Nielsen, and Hartl 2002; Hellmann et al. 2003; Keightley and Gaffney 2003; Urrutia and Hurst 2003; Chamary and Hurst 2004; Comeron 2004; Lavner and Kotlar 2005; Lu and Wu 2005). For example, constitutively and alternatively spliced exons differ in GC content at third (synonymous) sites (Iida and Akashi 2000).

What might be the mechanism for selection at so-called silent sites? The classical model, that selection favours efficient translation (e.g. Ikemura 1985; Bulmer, Wolfe, and Sharp 1991; Akashi and Eyre-Walker 1998; Duret 2002), may not apply in mammals (Duret 2002; dos Reis, Savva, and Wernisch 2004) (but see also Urrutia and Hurst 2003; Comeron 2004; Lavner and Kotlar 2005). Some evidence suggests that synonymous sites might be of importance in mRNA stability (Duan and Antezana

2003; Duan et al. 2003; Capon et al. 2004). Here we consider the possibility that purifying selection acts at synonymous sites in exons to ensure efficient pre-mRNA splicing (Willie and Majewski 2004; Chamary and Hurst 2005a).

Exons are defined by sequence located within introns, the 5' splice site, branch point and 3' splice site (Robberson, Cote, and Berget 1990). However, this tripartite signal (Fairbrother and Chasin 2000) is often necessary but not sufficient for intron excision. In human introns, these signals contain only half the required information for accurate splicing (Lim and Burge 2001). Exonic splicing enhancers (ESEs) are oligonucleotide sequences that are abundant in both constitutively and alternatively spliced exons (Tian and Kole 1995; Coulter, Landree, and Cooper 1997; Liu, Zhang, and Krainer 1998; Schaal and Maniatis 1999; Fairbrother et al. 2002). Most ESEs facilitate splice site recognition by recruiting serine-arginine-rich (SR) proteins during spliceosome assembly and localisation (Wang et al. 2004). The Burge/Sharp group recently developed a computational method (Fairbrother et al. 2002; Fairbrother et al. 2004b) that identifies candidate hexameric sequences with ESE activity. The density of ESE hexamers increases as one approaches intron-exon junctions (Fairbrother et al. 2004a) (Supplementary Figure 1). ESE activity is optimal within ~70 nucleotides of splice sites, although the effect is dependant on the strength of the enhancer, with potent enhancers exerting an influence at double this distance (Graveley, Hertel, and Maniatis 1998).

Prior evidence suggests that codon choice is biased owing to the presence of ESEs and biased against intronic splicing enhancers (Willie and Majewski 2004; Chamary and Hurst 2005a), e.g. the codon GAA is common in ESEs and is increasingly preferred over its synonym GAG near intron-exon boundaries. It is unclear, however, whether this explains all the trends in codon bias as a function of distance from exonic ends (Eskesen, Eskesen, and Ruvinsky 2004; Chamary and Hurst 2005a). Consistent with a preference for ESEs in certain exonic locations, at least two genes exhibit a marked reduction in the synonymous rate of evolution in regions containing an ESE (BRCA1: Hurst and Pal 2001; Liu et al. 2001; Orban and Olah 2001; CFTR: Pagani, Raponi, and Baralle 2005). More generally, it has been reported that single-nucleotide polymorphism (SNP) density decreases as one approaches the ends of exons (Majewski and Ott 2002) and that this can be explained by increasing ESE density (Fairbrother et al. 2004a). Although some ESEs appear to be conserved over the course of evolution (Yeo et al. 2004), it has not previously been demonstrated that certain mutations have been opposed by natural selection because they occur within ESEs. Consequently, here we ask whether putative ESEs are associated with a lower rate of synonymous evolution and, if they are, what impact this might have had on estimates of the mutation rate (μ) derived from the rate of synonymous nucleotide substitution (K_s).

Methods

Assemblage of the data sets

We downloaded the 7645 human-chimp-mouse orthologues used by Clark et al. (2003) from http://www.sciencemag.org/cgi/content/full/302/5652/1960/DC1, using only those alignments where all three sequences contained start and terminal stop codons. The GeneID (LocusLink) numbers in the annotation file were used to derive the human RefSeq identifiers at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene. We then compared the human sequences in the alignments to those in the RefSeq files, retaining only those which were the same length and >99% identical. The RefSeq identifier was then used to identify genomic sequence (hence exon structure of the human CDS) at EnsEMBL, http://www.ensembl.org/Homo_sapiens/exportview. We ignored EnsEMBL genomic files where the CDS of the associated RefSeq was not the same length as that derived from the genomic annotation. For the 973 genes remaining, the intron-exon junctions in the alignments were reconstructed from the genomic sequence.

Exonic splicing enhancer identification

Defining sequence as ESE or ESS is non-trivial, so we took several different approaches. In principle, a putative ESE within an alignment could be defined as sequence present in one, either or both species. Although one might imagine that the latter is the best definition because it is the most restrictive, human and mouse ESEs are very similar (e.g. 175/238 human hexamers are also found in mouse) and so this protocol may well end up isolating slow-evolving sequence, rather than ESE. Consider the following hypothetical human-mouse alignment:

Human GAAGAATTT

Mouse CCCGAAGAA

If the hexamer GAAGAA is only identified in one species (by 'human masking' or 'mouse masking'), 6 of the 9 sites are considered to be associated with the ESE (underlined) and 3 nucleotide substitutions have occurred. Under our most stringent definition of an ESE ('human+mouse masking'), only the 3 sites (GAA) that are within hexamers in both species are considered.

Evolutionary rate estimation

Non-synonymous (K_a) and synonymous (K_s) substitution rates were estimated with the Li method (Li 1993) using the Kimura 2-parameter model. To control for heterogeneity in mutation/substitution rates between genes (e.g. Lercher, Chamary, and Hurst 2004), differences in K between putative ESE and non-ESE were performed by paired analyses using one-sample Wilcoxon signed rank tests. To minimise the effect of noise when sampling short sequence, we only considered pairs of sequences (ESE

versus non-ESE) where neither rate estimate was unusually high for the comparison: human-chimp K_a<0.01 and K_s<0.03; human-mouse K_a<0.2 and K_s<0.75.

Proportion of substitutions as a function of distance from intron-exon junctions

Each exon was divided in two, with the first half being considered the 5' end and the second the 3' end. Under this protocol no given site can be counted more than once. Running towards the interior of an exon, the distance from the intron-exon junction is the number of nucleotides (including gaps) from the junction pertinent to the half-exon. If a given site was four-fold degenerate in both species, we incremented the count of the number of sites at that distance and the number of substitutions where appropriate.

Results

Synonymous evolution is slower in exonic splicing enhancers

If selection acts to preserve splicing activity (Yeo et al. 2004), the rate of synonymous substitution should be lower in putative ESEs when compared with non-ESE sequence. To investigate this we scanned a dataset of chimp-human-mouse orthologues (Clark et al. 2003) for the presence of 238 putative human (Fairbrother et al. 2002) and 380 mouse (Yeo et al. 2004) ESE hexamers. As ESEs have yet to be identified in chimp, here we report data for the human-mouse comparison, although the use of human hexamers as a 'chimp' set yields qualitatively the same results (Supplementary Table 1). Similarly, as many ESE are conserved (Yeo et al. 2004), one can also identify 'mammalian' enhancers. This too gives similar results (Supplementary Table 2).

As it is unclear on a priori grounds whether we should consider putative ESEs as being present in one or both species, we employ various masking protocols to identify sites that might be associated with putative ESEs. The first method identifies ESE sites as those that occur within human hexamers in human sequence ('human masking'). The second considers ESE sites to be those that are within mouse hexamers ('mouse masking'). Using more stringent definitions, we can also define ESE sites to be those present within hexamers in both sequences ('human+mouse masking').

In all masking permutations, we find that the synonymous substitution rate in putative ESEs is lower than that in non-ESEs (Table 1; Supplementary Table 1). The magnitude of the reduction in K_s is dependent on the masking protocol. The difference in K_s is relatively modest when masking hexamers in single species (~5%) but quite large in the more stringent double masking (~35%).

The reduction in K_s within ESEs is not due to a skewed CpG distribution

Sites within CpG dinucleotides are known to be hypermutable (Bird 1980; Cooper and Krawczak 1989; Sved and Bird 1990) and ESEs are typically purine-rich (Blencowe 2000) (in combined human/mouse hexamers A=42.5%, G=25.7% C=17.9%, T=13.9%). Consequently, it is possible that the reduction in K_s is an artefact owing to non-ESE sequence having a higher concentration of CpGs. However, after repeating the above analysis, this time omitting CG/GC pairs in either sequence, we again find that putative ESEs evolve more slowly than non-ESEs (Table 1). In fact, the previously marginally non-significant difference in the mouse masking now becomes significant. We conclude that the decreased K_s in ESEs cannot be explained by differential abundances of hypermutable CpGs.

The reduction in K_s within ESEs is not due to a skewed nucleotide distribution

The above test considers a class of well-known hypermutable sites. However, different nucleotides may themselves have different mutabilities (see e.g. Chamary and Hurst 2004). More generally we can ask whether, controlling for skewed nucleotide contents, ESEs still have unusually low synonymous rates of evolution. Moreover, it is also possible that the reduced K_s is a result of searching for relatively little sequence (particularly in human+mouse masking) which will artificially isolate slowly evolving sequences.

To examine these possibilities we performed a simulation. In each of 1000 randomisations we generated a set of simulated hexamers of the same average nucleotide composition as the real ESE hexamers. These simulated sets are then used to carry out human, mouse and the human+mouse (stringent) maskings. For each gene the difference between the real and the simulants was expressed by a Z-score, the number of standard deviations the observed K_s (from real ESEs) is away from the mean K_s of the simulated ESEs. Under a null hypothesis that the reduced K_s in ESE is due to the masking protocol and/or skewed nucleotide content in ESEs, the Z-score distribution should have a mean that is not significantly different to zero. Alternatively, if putative ESEs evolve slowly, then their K_s should be significantly lower than the mean of the simulants, i.e. a negative Z-score. Under the three protocols studied, we found that this was indeed the case (human masking median Z=-0.293, P<0.0001; mouse median Z=-0.214, P<0.0001; human+mouse median Z=-0.17, P=0.015). We conclude that the low K_s in putative ESEs is not owing to skewed nucleotide content or any bias introduced by the masking process.

A decline in substitution rate at four-fold degenerate sites near intron-exon junctions is partially explained by the presence of ESEs

While the above results are consistent with a model in which ESE sequence is under selection to retain their function, there exists a further possibility. ESE density is known to be highest near intron–exon junctions. If, for some other reason, ESEs in the near vicinity of such junctions are under stronger selection (or experience low mutation rates), then they would have lower rates of evolution than either non-ESE sequence or our simulated-ESEs, both of which may be relatively more common in exonic interiors. For example, exon-exon junctions tend to occur at or around the position of nucleosome formation (Kogan and Trifonov 2005). If nucleosomal or peri-nucleosomal sequence is more conserved than the average, then we may expect ESEs to be slow-evolving, but only because they tend to be near nucleosomes. Note too that there may well be patterns of nucleotide usage across exons that are not explained by ESE presence/absence (Eskesen, Eskesen, and Ruvinsky 2004; Chamary and Hurst 2005a). We can therefore ask whether, given their location in proximity to the junctions, ESEs evolve slower than non-ESEs and whether this alone is adequate to explain the reduced SNP density near intron-exon junctions (Fairbrother et al. 2004a).

The frequency of substitutions at four-fold degenerate sites was assessed as a function of distance from both 5' and 3' ends of exons (i.e. without masking ESE/non-ESE, but ignoring CpGs). This strongly suggests that synonymous mutations are increasingly opposed as one approaches the end of an exon (Figure 1). Studies looking at SNP density have suggested that such selection only extends about 30 nucleotides into exons (Majewski and Ott 2002; Fairbrother et al. 2004a), but we observe an effect that is closer to the biased codon choice data (~100 nucleotides, Willie and Majewski 2004; Chamary and Hurst 2005a).

Given the possible discrepancy in the scale of the effect, we then asked whether the effect is likely owing to a reduced rate of evolution in ESEs coupled with their greater proximity to intron-exon junctions, or to some more general underlying cause. Under the first model, we expect ESE rates of evolution and non-ESE rates of evolution both to show no trend as a function of the distance from the junction, but with the ESE synonymous rates lower than those of the non-ESEs. In the second case, we might expect ESE and non-ESE to show the same trend of increasing synonymous divergence as a function of distance from the junction and no difference in the rates of evolution controlling for distance from junction.

These hypotheses were tested by analysis of covariance (ANCOVA) in which the distance from the junction was the covariat and enhancer and non-enhancer sequence were the two factors/groups (NB there is no significant interaction term so the assumptions of ANCOVA are upheld). The difference in rates between the groups was always significant controlling for the distance from the junction. This strongly suggests

that ESEs are slow evolving even controlling for their differential abundance near junctions (Table 2; Figure 2). In all cases, there remains an effect whereby all sequences evolve marginally slower if closer to the junction. This suggests the presence of some weak force affecting substitution rates as a function of the distance from the boundary independent of ESE presence or absence.

The effect of ESEs on evolution at non-synonymous sites

Here we have concentrated on how conservation of ESEs can influence synonymous mutations and codon usage. In principle, however, ESEs could also affect non-synonymous mutations. This may well be the case, as K_a is lower in putative ESEs (Table 3). Moreover, it is interesting to ask whether amino acids specified by purine-rich codons are also more abundant near junctions. If so we should expect the effect to be most strikingly seen for usage of lysine (AAA and AAG), A being the most common nucleotide in ESEs followed by G. This is indeed observed (Figure 3). However, while AG rich codons tend to be employed near boundaries, at least for the 3' end, the effect is more striking for AT rich codons (see Supplementary Figure 2). This suggests a pressure towards A and T rather than A and G and might hint at some other force. This is unlikely to be nucleosome associated as in mouse and human these are associated with G and C (Kogan and Trifonov 2005).

Discussion

From the above analyses it appears that splicing enhancers are under purifying selection. As the enhancer regions do not discriminate synonymous from non-synonymous sites, this is reflected in a reduced synonymous rate of evolution within these domains. Assuming this to be the only form of selection on silent sites, it is pertinent to ask how much we might have underestimated the real mutation rate by employing rates of synonymous evolution. To address this issue we need to know what proportion of the sequence is functional splicing enhancer and what, on the average is the reduction in the rate of evolution of functional splicing enhancer.

We have employed three different methods to define putative splicing enhancer. Putative enhancers identified within a single species (mouse or human) show a modest 1%-11% reduction in their rate of evolution (depending on whether we ignore CpGs). Those parts that are enhancer in both mouse and human have a more striking ~35% reduction in their rate compared with non-enhancer regions. However, the more stringent definition defines less of the sequence as being in enhancer. When we factor in the proportion of sequence that is putatively enhancer, the three methods all suggest the same answer, namely that the net reduction in Ks, owing to the presence of splice enhancers, is modest. It may be as low as 2% and unlikely to be much more than 8% (Table 4). This suggests that correction for the presence of ESEs will not have a major

effect on estimates of the mutation rate, not least because the margin of error associated with estimates of the number of generations between any two mammalian taxa is vastly more error-prone and alterations here will have a much more profound effect.

Conservation of ESEs is unlikely to be the only form of selection at synonymous sites. For splicing-associated functions, for example, we have not considered the contribution of exonic splicing silencer (ESS) sequence, although we find that masking the 133 decamers that have been systematically identified in humans (Wang et al. 2004) does not alter our conclusions (Supplementary Table 3). Biased codon usage may also reflect an avoidance of certain sequences (Eskesen, Eskesen, and Ruvinsky 2004; Chamary and Hurst 2005a). Importantly, the strongest signal for selection that has been seen so far is a high stability of cytosine at third sites (Chamary and Hurst 2004). This is not obviously explained by a role in the splicing process (Chamary and Hurst 2005a) because ESEs are AG-rich and C-poor. Therefore we cannot conclude that selection on silent sites has not lead to a significant underestimate of the mutation rate. The cause of the C preference remains unclear, but a role in mRNA stability is supported by some data (Chamary and Hurst 2005b).

One consequence of all this evidence for skewed nucleotide composition and biased codon usage near intron-exon boundaries is that it adds layers of complexity to the interpretation of prior results. For example, several recent reports find evidence for systematic codon bias that is not explained by background nucleotide content (Urrutia and Hurst 2003; Comeron 2004; Lavner and Kotlar 2005). For example, highly expressed genes exhibit the greatest bias (Urrutia and Hurst 2003). From the above results, we might expect some form of systematic variation with intron density. As this also varies with expression parameters (Comeron 2004), one may simply be observing a higher proportion of codons in the vicinity of intron-exon boundaries. Clearly, to safely conclude that a bias exists, one should exclude those regions of exons within about 70 nucleotides either side of junctions.

The putative impact of ESEs on non-synonymous substitution rates has numerous corollaries. First, this makes it difficult to ask whether a certain protein domain is under purifying selection. A low K_a may be evidence for this, but it could also be explained by selection on the enhancer rather than the protein. To examine in detail such claims, one should also ask whether the DNA specifying the domain is near an intron-exon junction and matches known ESEs. The skewed amino acid usage near intron-exon boundaries has two possible interpretations. First, that at the time of insertion, a viable intron can only be tolerated if there are already ESEs present in the near vicinity. Second, that after insertion, the process of splicing is subject to selection, with choice of amino acids around junctions being determined in part by the efficiency of splicing of flanking introns. These are not mutually incompatible. To establish

whether the first is true, one would need to identify new introns within the mammal lineage. These are, to the best of our knowledge, remarkably rare (but see Sry in marsupials, O'Neill et al. 1998).

References

Akashi, H., and A. Eyre-Walker. 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. **8**:688-693.

Bird, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. **8**:1499-1504.

Blencowe, B. J. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem. Sci. **25**:106-110.

Bulmer, M., K. H. Wolfe, and P. M. Sharp. 1991. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc. Natl Acad. Sci. USA **88**:5974-5978.

Bustamante, C. D., R. Nielsen, and D. L. Hartl. 2002. A maximum likelihood method for analyzing pseudogene evolution: Implications for silent site evolution in humans and rodents. Mol. Biol. Evol. **19**:110-117.

Capon, F., M. H. Allen, M. Ameen, A. D. Burden, D. Tillman, J. N. Barker, and R. C. Trembath. 2004. A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. Hum. Mol. Genet. **13**:2361-2368.

Chamary, J. V., and L. D. Hurst. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: Evidence for selectively driven codon usage. Mol. Biol. Evol. **21**:1014-1023.

Chamary, J. V., and L. D. Hurst. 2005a. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet. **21**:256-259.

Chamary, J. V., and L. D. Hurst. 2005b. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. **6**:R75.

Clark, A. G., S. Glanowski, R. Nielsen et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302**:1960-1963.

Comeron, J. M. 2004. Selective and mutational patterns associated with gene expression in humans: Influences on synonymous composition and intron presence. Genetics **167**:1293-1304.

Cooper, D. N., and M. Krawczak. 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. **83**:181-188.

Coulter, L. R., M. A. Landree, and T. A. Cooper. 1997. Identification of a new class of exonic splicing enhancers by in vivo selection. Mol. Cell. Biol. **17**:2143-2150.

dos Reis, M., R. Savva, and L. Wernisch. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. **32**:5036-5044.

Duan, J., and M. A. Antezana. 2003. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J. Mol. Evol. **57**:694-701.

Duan, J., M. S. Wainwright, J. M. Comeron, N. Saitou, A. R. Sanders, J. Gelernter, and P. V. Gejman. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum. Mol. Genet. **12**:205-216.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12**:640-649.

Eskesen, S. T., F. N. Eskesen, and A. Ruvinsky. 2004. Natural selection affects frequencies of AG and GT dinucleotides at the 5' and 3' ends of exons. Genetics **167**:543-550.

Eyre-Walker, A. 1991. An analysis of codon usage in mammals: selection or mutation bias? J. Mol. Evol. **33**:442-449.

Eyre-Walker, A., and P. D. Keightley. 1999. High genomic deleterious mutation rates in hominids. Nature **397**:344-347.

Fairbrother, W. G., and L. A. Chasin. 2000. Human genomic sequences that inhibit splicing. Mol. Cell. Biol. **20**:6816-6825.

Fairbrother, W. G., D. Holste, C. B. Burge, and P. A. Sharp. 2004a. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. **2**:e268.

Fairbrother, W. G., R. F. Yeh, P. A. Sharp, and C. B. Burge. 2002. Predictive identification of exonic splicing enhancers in human genes. Science **297**:1007-1013.

Fairbrother, W. G., G. W. Yeo, R. Yeh, P. Goldstein, M. Mawson, P. A. Sharp, and C. B. Burge. 2004b. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res. **32**:W187-190.

Graveley, B. R., K. J. Hertel, and T. Maniatis. 1998. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. EMBO J. **17**:6747-6756.

Hellmann, I., S. Zollner, W. Enard, I. Ebersberger, B. Nickel, and S. Paabo. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. **13**:831-837.

Hurst, L. D., and C. Pal. 2001. Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet. **17**:62-65.

Iida, K., and H. Akashi. 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene **261**:93-105.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13-34.

Keightley, P. D., and A. Eyre-Walker. 2000. Deleterious mutations and the evolution of sex. Science **290**:331-333.

Keightley, P. D., and D. J. Gaffney. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proc. Natl Acad. Sci. USA **100**:13402-13406.

Kimura, M. 1983. The Neutral Theory of Evolution. Cambridge University Press, Cambridge.

Kogan, S., and E. N. Trifonov. 2005. Gene splice sites correlate with nucleosome positions. Gene **352**:57-62.

Lavner, Y., and D. Kotlar. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene **345**:127-138.

Lercher, M. J., J. V. Chamary, and L. D. Hurst. 2004. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. Genome Res. **14**:1002-1013.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**:96-99.

Lim, L. P., and C. B. Burge. 2001. A computational analysis of sequence features involved in recognition of short introns. Proc. Natl Acad. Sci. USA **98**:11193-11198.

Liu, H. X., L. Cartegni, M. Q. Zhang, and A. R. Krainer. 2001. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. Nat. Genet. **27**:55-58.

Liu, H. X., M. Zhang, and A. R. Krainer. 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev. **12**:1998-2012.

Lu, J., and C. I. Wu. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc. Natl Acad. Sci. USA **102**:4063-4067.

Majewski, J., and J. Ott. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res. **12**:1827-1836.

O'Neill, R. J., F. E. Brennan, M. L. Delbridge, R. H. Crozier, and J. A. Graves. 1998. De novo insertion of an intron into the mammalian sex determining gene, SRY. Proc. Natl Acad. Sci. USA **95**:1653-1657.

Orban, T. I., and E. Olah. 2001. Purifying selection on silent sites - a constraint from splicing regulation? Trends Genet. **17**:252-253.

Pagani, F., M. Raponi, and F. E. Baralle. 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc. Natl Acad. Sci. USA **102**:6368-6372.

Robberson, B. L., G. J. Cote, and S. M. Berget. 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. **10**:84-94.

Schaal, T. D., and T. Maniatis. 1999. Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol. Cell. Biol. **19**:1705-1719.

Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi, and J. F. Peden. 1995. DNA-sequence evolution: the sounds of silence. Phil. Trans. R. Soc. Lond. B **349**:241-247.

Sved, J., and A. Bird. 1990. The Expected Equilibrium of the CpG Dinucleotide in Vertebrate Genomes Under a Mutation Model. Proc. Natl Acad. Sci. USA **87**:4692-4696.

Tian, H., and R. Kole. 1995. Selection of novel exon recognition elements from a pool of random sequences. Mol. Cell. Biol. **15**:6291-6298.

Urrutia, A. O., and L. D. Hurst. 2003. The signature of selection mediated by expression on human genes. Genome Res. **13**:2260-2264.

Wang, Z., M. E. Rolish, G. Yeo, V. Tung, M. Mawson, and C. B. Burge. 2004. Systematic identification and analysis of exonic splicing silencers. Cell **119**:831-845.

Willie, E., and J. Majewski. 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet. **20**:534-538.

Yeo, G., S. Hoon, B. Venkatesh, and C. B. Burge. 2004. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc. Natl Acad. Sci. USA **101**:15700-15705.

Table 1. Differences in the rate of synonymous evolution between putative ESE and non-ESE sequence in human-mouse alignments

Masking protocol[a]	Non-ESE[b]	ESE[b]	N^c	P^d
Human	0.4484 ± 0.0042	0.4117 ± 0.0054	812	8e-11
Human non-CpG	0.3378 ± 0.0041	0.3006 ± 0.0053	848	1e-12
Mouse	0.4440 ± 0.0040	0.4377 ± 0.0048	854	0.0538
Mouse non-CpG	0.3343 ± 0.0041	0.3184 ± 0.0048	889	8e-05
Human+mouse	0.4701 ± 0.0042	0.2896 ± 0.0053	815	3e-103
Human+mouse non-CpG	0.3488 ± 0.0041	0.2157 ± 0.0048	797	3e-77

[a] The sequences in which putative exonic splicing enhancer (ESE) motifs are masked. For human+mouse, these are the sites that are identified as being associated with ESEs in both species.

[b] The mean synonymous substitution rate (± SEM).

[c] The number of genes analysed in pairwise comparisons.

[d] The significance of the difference between ESE and non-ESE (*P*-values from paired t-tests).

Table 2. Analysis of covariance between putative ESE and non-ESE sequence for the substitution frequency at four-fold synonymous sites as a function of distance from intron-exon junctions in human-mouse alignments

Masking protocol[a]	Parameter	5' end of exons		3' end of exons	
		Estimate[b]	P^c	Estimate[b]	P^c
Human non-CpG	Distance	0.0005 ± 0.0001	7e-05	0.0003 ± 0.0001	0.0137
	Level	0.0254 ± 0.0054	7e-06	0.0214 ± 0.0060	0.0005
Mouse non-CpG	Distance	0.0005 ± 0.0001	0.0002	0.0003 ± 0.0001	0.0123
	Level	0.0231 ± 0.0053	3e-05	0.0376 ± 0.0051	2e-11
Human+mouse non-CpG	Distance	0.0005 ± 0.0001	0.0001	0.0003 ± 0.0001	0.0244
	Level	0.0886 ± 0.0061	<2e-16	0.1036 ± 0.0067	<2e-16

[a] The sequences in which putative ESEs are masked.

[b] The 'Estimate' for 'Distance' is the slope of the regression line (± SEM) for the substitution frequency at four-fold sites in ESEs plotted against the distance from the intron-exon junction. There is no difference between the slopes derived from ESE and non-ESE sequence ($P > 0.05$). The 'Estimate' for 'Level' is the difference between the slopes (± SEM) for ESE and non-ESE.

[c] For 'Distance', the *P*-value indicates whether the common slope (ESE was used) is significant. For 'Level', the *P*-value indicates whether there is a difference between ESEs and non-ESEs while controlling for the distance from the junction, i.e. to determine whether, at a given distance from the junction, the proportion of substitutions at four-fold sites differs between ESE and non-ESE.

Table 3. Differences in the rate of amino acid evolution between putative ESE and non-ESE sequence in human-mouse alignments

Masking protocol[a]	Non-ESE[b]	ESE[b]	N^c	P^d
Human	0.0526 ± 0.0015	0.0473 ± 0.0015	862	5e-09
Human non-CpG	0.0394 ± 0.0013	0.0404 ± 0.0015	874	0.5685
Mouse	0.0524 ± 0.0015	0.0503 ± 0.0015	890	0.0147
Mouse non-CpG	0.0396 ± 0.0013	0.0402 ± 0.0014	908	0.4211
Human+mouse	0.0545 ± 0.0016	0.0343 ± 0.0013	838	2e-68
Human+mouse non-CpG	0.0418 ± 0.0015	0.0298 ± 0.0013	815	1e-34

[a] The sequences in which putative ESEs are masked.

[b] The mean non-synonymous substitution rate (± SEM).

[c] The number of genes analysed in pairwise comparisons.

[d] The significance of the difference between ESE and non-ESE (*P*-values from paired t-tests).

Table 4. The contribution of purifying selection at synonymous sites in putative ESEs to underestimates of the mutation rate (μ) in mammals

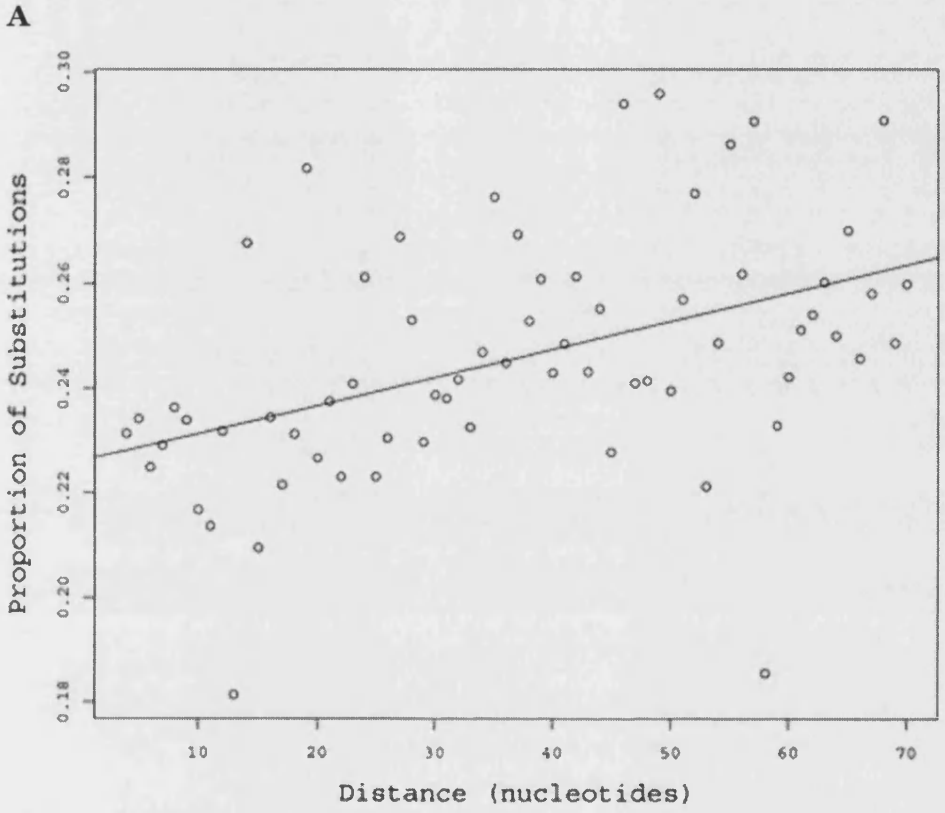
Masking protocol[a]	K_s reduction[b] (%)	ESE coverage[c] (%)	μ underestimation (%)
Human	8.19	30.42	2.49
Human non-CpG	11.03	30.42	3.36
Mouse	1.41	40.30	0.57
Mouse non-CpG	4.74	40.30	1.91
Human+mouse	38.39	21.77	8.36
Human+mouse non-CpG	38.15	21.77	8.31

[a] The sequences in which putative ESEs are masked.

[b] The difference in the synonymous substitution rate between ESE and non-ESE.

[c] The proportion of sequence covered by ESE sites.

Figure 1. Frequency of substitutions at four-fold degenerate sites in human-mouse alignments as a function of distance from intron-exon junctions, at (A) the 5' end of exons (slope = 0.2260; R^2 = 0.1995; P = 9e-05) and (B) the 3' end (slope = 0.2372; R^2 = 0.0660; P = 0.0203). The lines of best fit are derived by linear regression and weighted by the number of sites.



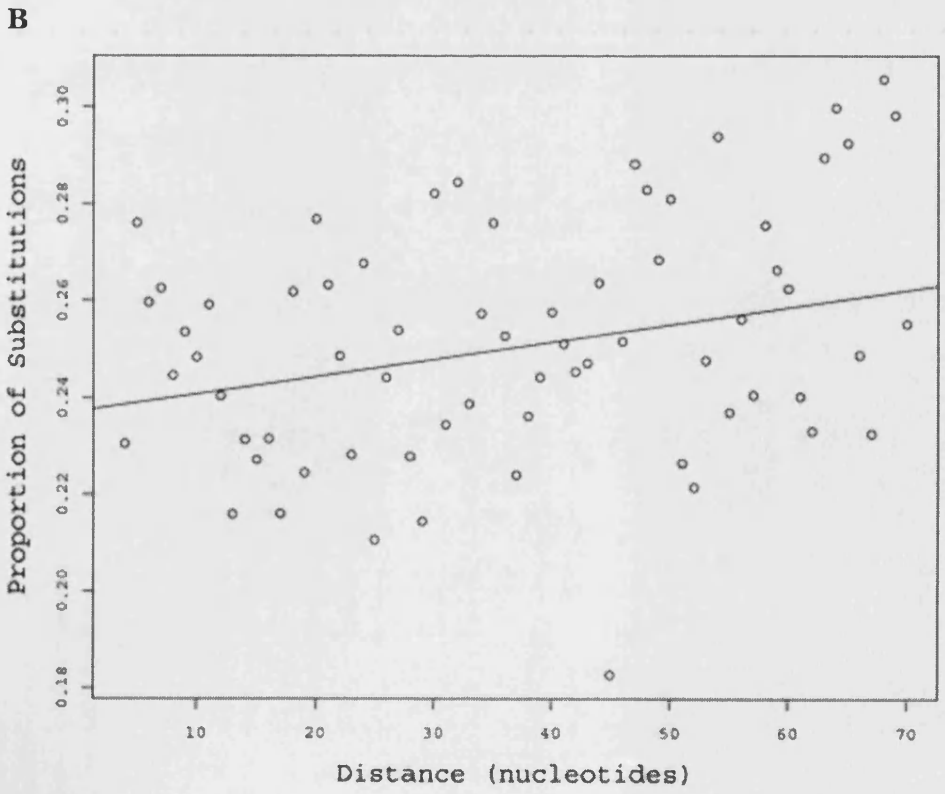
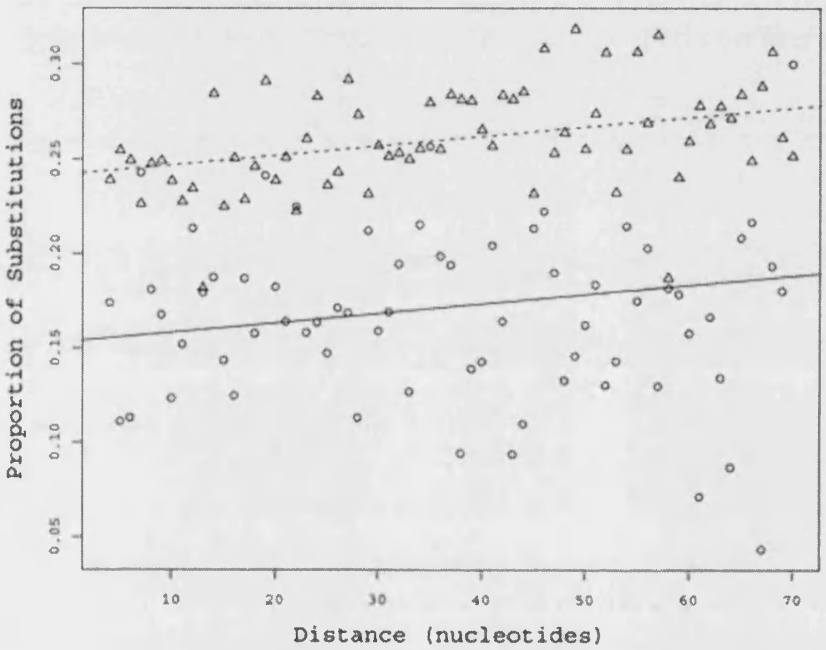


Figure 2. Frequency of substitutions at four-fold degenerate sites in human-mouse alignments as a function of distance from intron-exon junctions in ESE (circles and solid lines) and non-ESE (triangles and dashed lines) sequence, at (A) the 5' end of exons and (B) the 3' end. The weak trends are shown for sites within ESEs at the 5' (A, slope = 0.1658; R^2 = 0; P = 0.6831) and 3' end (B, slope = 0.1369; R^2 = 0.0773; P = 0.0130), and non- ESE sequence at the 5' (A, slope = 0.2396; R^2 = 0.1625; P = 0.0004) and 3' end (B, slope = 0.2575; R^2 = 0.0143; P = 0.1664). The lines of best fit are derived by linear regression and weighted by the number of nucleotide sites. The ESE masking is by the human+mouse protocol.

A



B

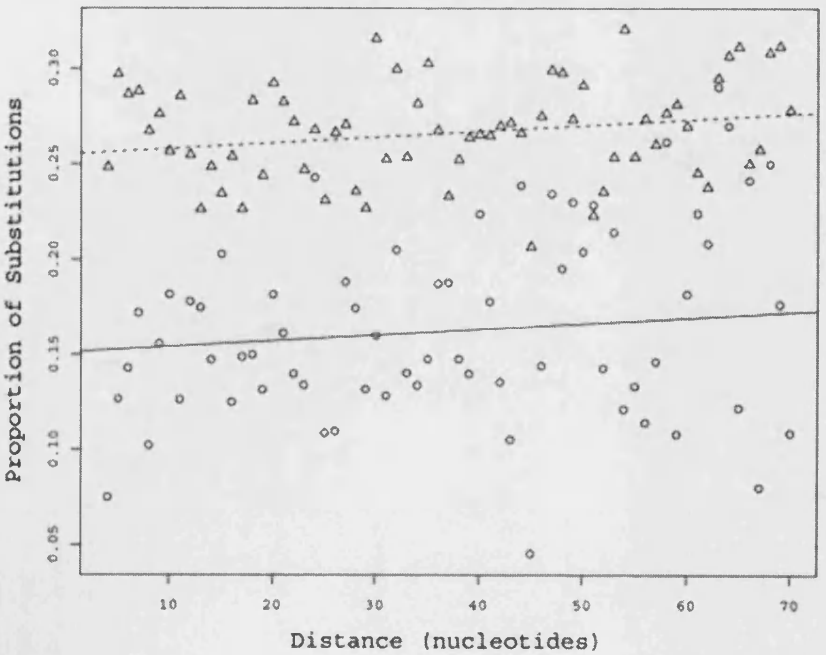
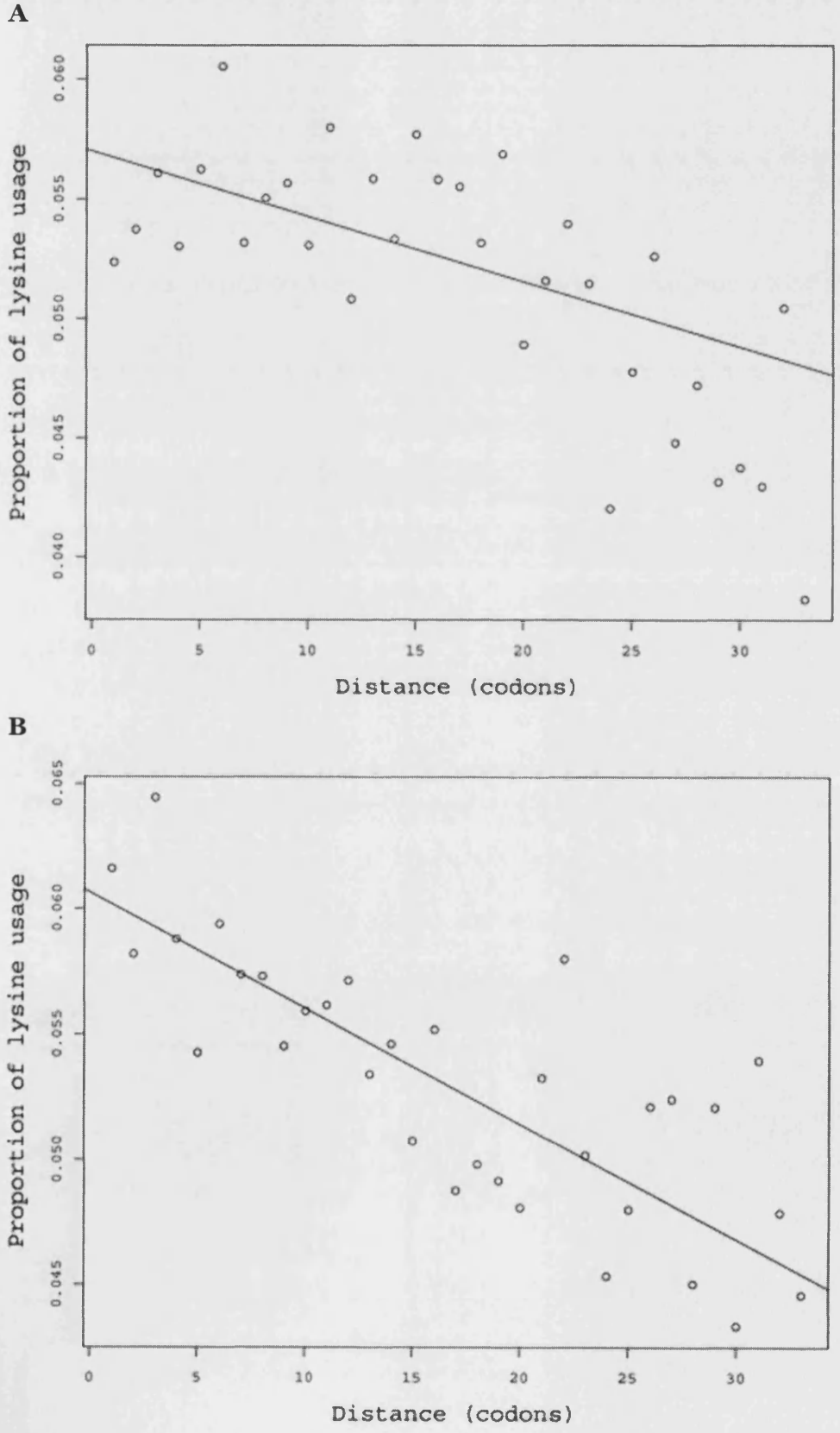


Figure 3. Lysine residue usage as a function of distance from intron-exon junctions, at (A) the 5' end of exons (R^2 = 0.2936; P = 0.0007) and (B) the 3' end (R^2 = 0.6364; P = 2e-08). The lines of best fit are derived by linear regression and weighted by the number of codons.



Supplementary Table 1. Differences in the rate of synonymous evolution between putative ESE and non-ESE sequence in human-chimp alignments

Masking protocol[a]	Non-ESE[b]	ESE[b]	N^c	P^d
Human	0.0091 ± 0.0002	0.0046 ± 0.0002	803	1e-30
Human non-CpG	0.0044 ± 0.0002	0.0023 ± 0.0002	846	1e-12
Chimp	0.0092 ± 0.0003	0.0045 ± 0.0003	788	9e-33
Chimp non-CpG	0.0043 ± 0.0002	0.0024 ± 0.0001	842	3e-11
Human+chimp	0.0103 ± 0.0003	0.0020 ± 0.0002	819	2e-82
Human+chimp non-CpG	0.0050 ± 0.0002	0.0015 ± 0.0002	863	2e-31

[a] The sequences in which putative exonic splicing enhancer (ESE) motifs are masked. For human+mouse, these are the sites that are identified as being associated with ESEs in both species. Note that ESEs for chimp have not been defined, so human hexamers were used as a 'chimp' ESE set to mask chimp sequence.

[b] The mean synonymous substitution rate (± SEM).

[c] The number of genes analysed in pairwise comparisons.

[d] The significance of the difference between ESE and non-ESE (*P*-values from paired t-tests).

Supplementary Table 2. Differences in the rate of synonymous evolution between putative mammalian ESE and non-ESE sequence in human-mouse alignments

Masking protocol[a]	Non-ESE[b]	ESE[b]	N^c	P^d
Combined	0.4971 ± 0.0041	0.2723 ± 0.0043	881	2e-134
Combined non-CpG	0.3791 ± 0.0042	0.2007 ± 0.0042	897	6e-114
Conserved	0.4786 ± 0.0041	0.1957 ± 0.0044	801	7e-128
Conserved non-CpG	0.3561 ± 0.0041	0.1664 ± 0.0044	771	2e-102

[a] The sequences in which putative ESEs are masked. 'Combined' is a set containing all 443 human and mouse hexamers. 'Conserved' is a set containing 175 hexamers common to both the human and mouse sets. Sites are designated part of an ESE if present as a hexamer in both sequences (i.e. human+mouse masking).

[b] The mean synonymous substitution rate (± SEM).

[c] The number of genes analysed in pairwise comparisons.

[d] The significance of the difference between ESE and non-ESE (*P*-values from paired t-tests).

Supplementary Table 3. Differences in the rate of synonymous evolution between putative ESE/ESS and non-ESE/ESS sequence in human-mouse alignments

Masking protocol[a]	Non-ESE/ESS[b]	ESE/ESS[b]	*N*[c]	*P*[d]
Human	0.4486 ± 0.0042	0.4120 ± 0.0054	814	1e-10
Human non-CpG	0.3377 ± 0.0042	0.3006 ± 0.0053	847	2e-12
Human+mouse	0.4701 ± 0.0042	0.2896 ± 0.0054	815	3e-103
Human+mouse non-CpG	0.3488 ± 0.0041	0.2157 ± 0.0049	797	3e-77

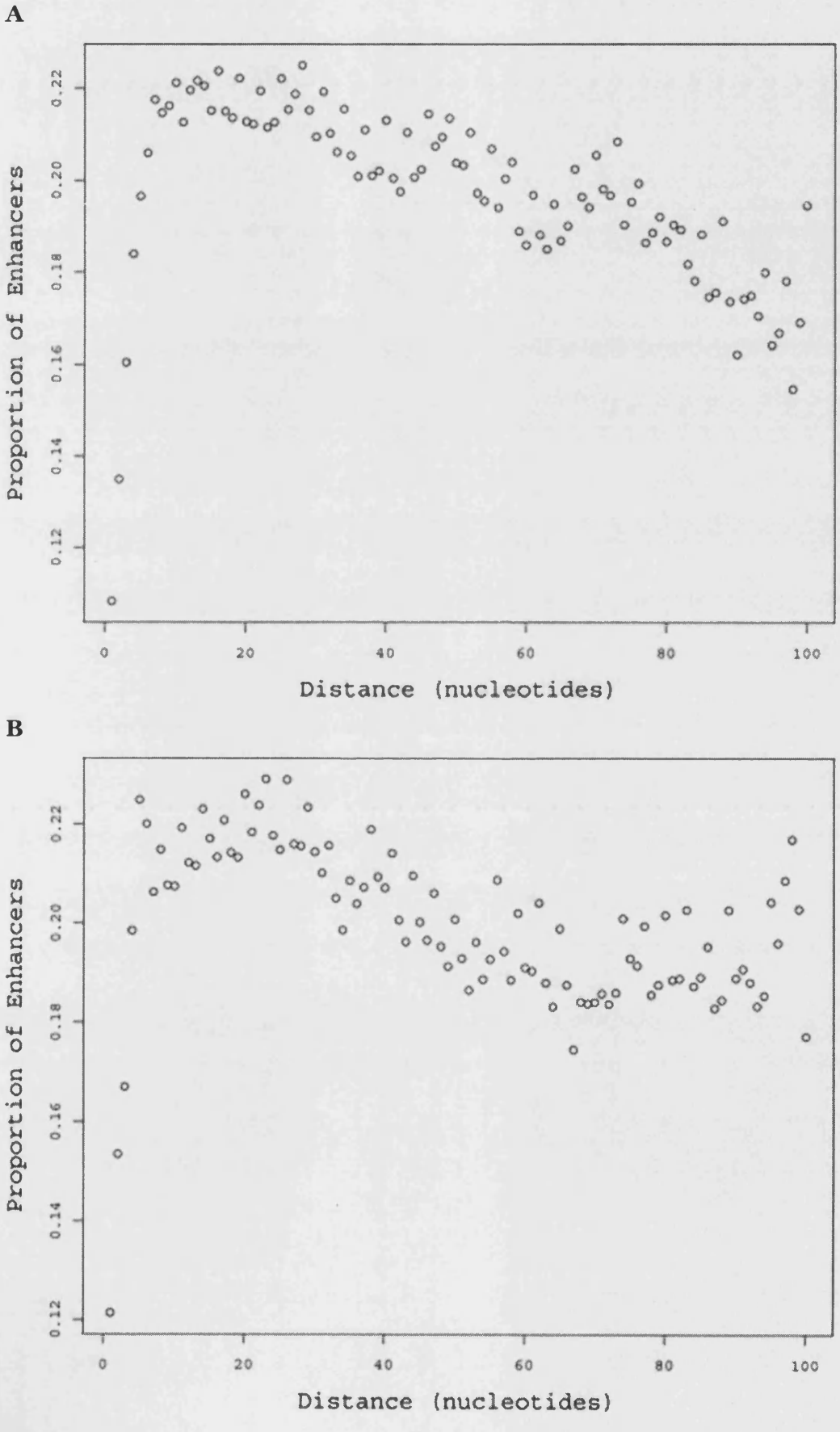
[a] The sequences in which putative ESE and exonic splicing silencer (ESS) motifs are masked. ESS decamers (133) have only been identified for human.

[b] The mean synonymous substitution rate (± SEM) in ESE/ESS and non-ESE/ESS sequence.

[c] The number of genes analysed in pairwise comparisons.

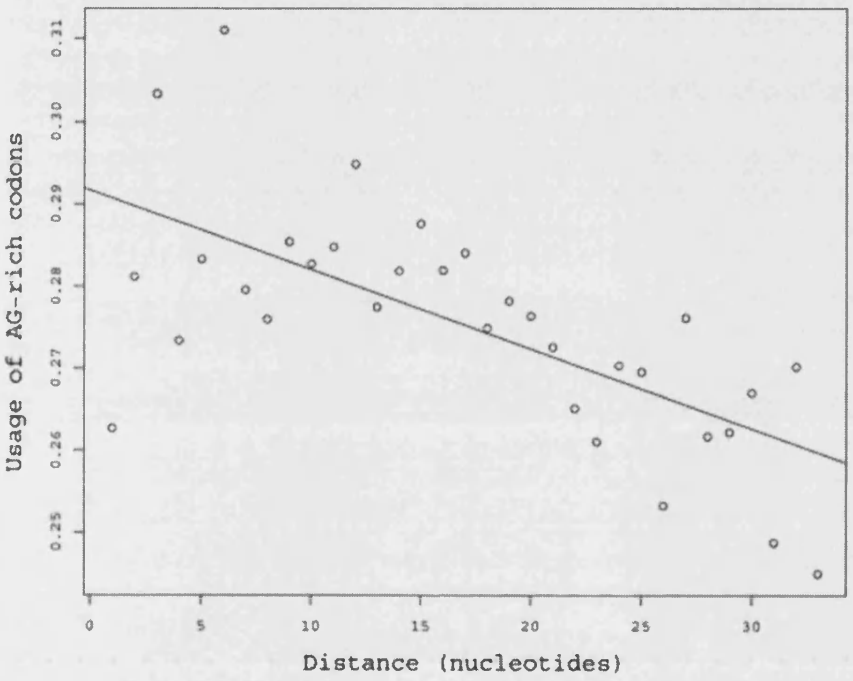
[d] The significance of the difference between putative ESE/ESS and non-ESE/ESS (*P*-values from paired t-tests).

Supplementary Figure 1. Frequency of ESE sites as a function of distance from intron-exon junctions, at (A) the 5' end of exons and (B) the 3' end.

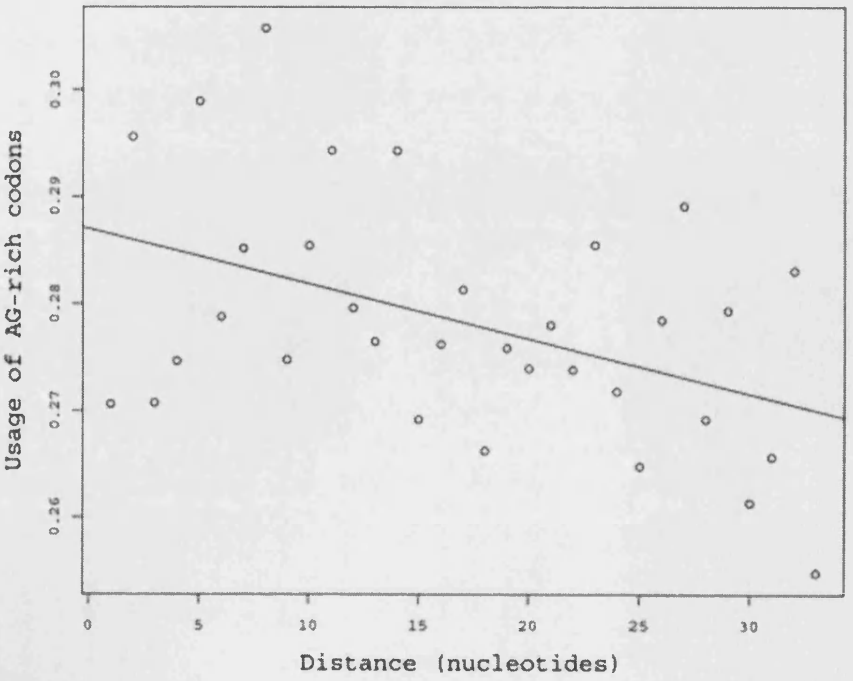


Supplementary Figure 2. Usage of AG-rich and AT-rich codons as a function of distance from intron-exon boundaries. Usage of codons that are (A) AG-rich at the 5' end of exons (R^2 = 0.432; P = 2e-05), (B) the AG-rich at the 3' end (R^2 = 0.1716; P = 0.0096), (C) AT-rich at the 5' end (R^2 = 0.549; P = 5e-07) and (D) AT-rich at the 3' end (R^2 = 0.709; P = 5e-10). The lines of best fit are derived by linear regression and weighted by the number of codons.

A

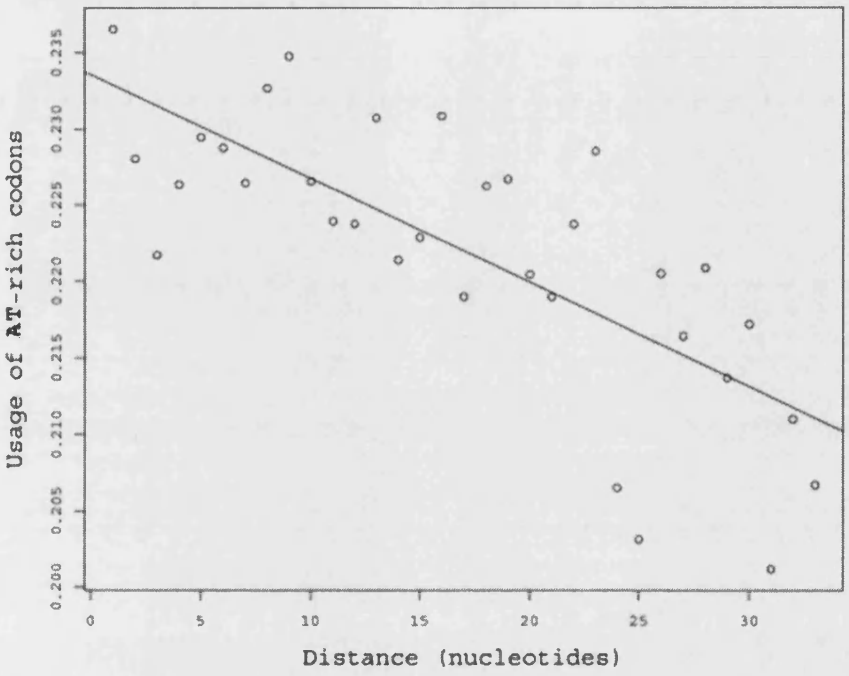


B

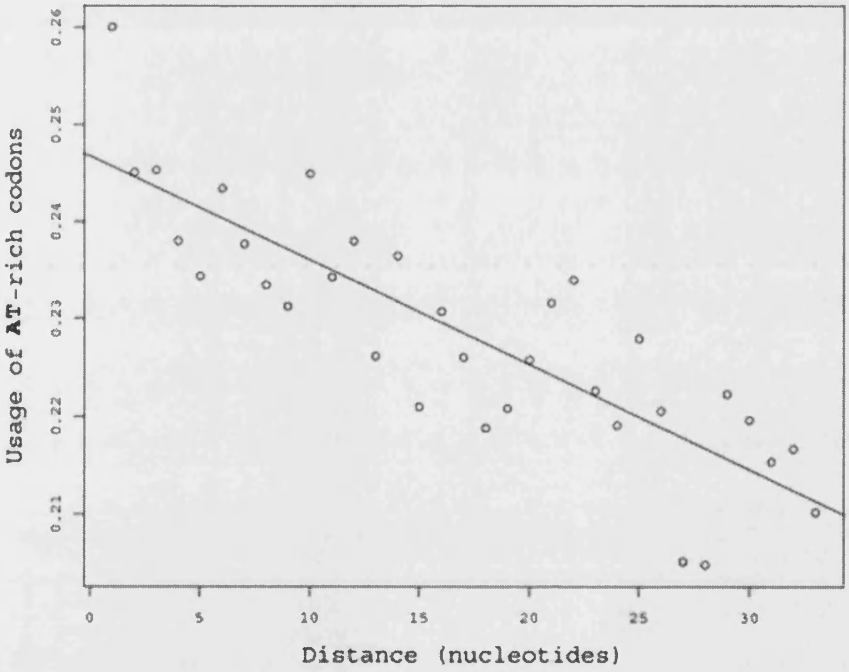


Supplementary Figure 2. (Continued)

C



D



Chapter 7. **Hearing silence: non-neutral evolution at synonymous sites in mammals**

Jean-Vincent Chamary, Joanna L. Parmley & Laurence D. Hurst

Nature Reviews Genetics (submitted)

The presumption of the neutral theory of molecular evolution, that some classes of mutation must be of too small an effect on fitness to be affected by natural selection, seems intuitively reasonable. Nonetheless, over the last few decades the neutral theory has been in retreat. At least in species with large populations, even synonymous mutations are not neutral. In mammals, by contrast, neutrality of these mutations is still commonly assumed. Here, however, we review evidence suggesting that even synonymous mutations are subject to constraint, often by affecting splicing and/or mRNA stability. This has implications for understanding disease, optimising transgenes, detecting positive selection and estimating the mutation rate.

Since its formulation in the 1960s, the neutral theory (BOX 1) has functioned as a powerful null model for molecular evolution[1]. The unexpectedly high rate of evolution of genes suggested that most mutations have no impact on an organism's fitness and so spread to fixation by chance[2] (drift). If all the mutations in putatively neutrally evolving DNA (e.g. introns, intergene spacer, synonymous sites) really are neutrally evolving, then the rate of evolution of such sequence can be used as a convenient measure of the mutation rate (e.g. REFS. 3-6). This does not require that all such mutations have absolutely no fitness consequence, just that they must be of such a small effect that the evolution of the mutations occurs as if they were neutral (BOX 1). For an allele to be 'effectively neutral', the selective disadvantage associated with it must be considerably smaller than the inverse of the effective population size (BOX 1). Consequently, we should expect neutral or effectively neutral evolution to be more common in species with small populations.

Even non-coding DNA (introns and intergene spacer), however, contains conserved functional elements. Is there then any class of sequence that we can confidently assume evolves neutrally and hence from which we can derive accurate estimates of the mutation rate? Taking a historical view, we note that synonymous mutations in mammals are almost unique in still being considered neutral. This is because mammals are unusual in having small populations (so rendering mutations of slight effect effectively neutral) and because codon usage appears to be largely dictated by patterns of base composition in the genomic region (isochore) within which a gene resides. We then propose, however, that this position requires substantial revision, noting that recent evidence suggests synonymous sites are important, for example, in mRNA stability and for correct splicing.

The rise and fall of the neutral theory

The original neutral theory proposed that both mutations having no effect on amino acid content (non-coding or synonymous changes) and those altering proteins (non-synonymous changes), could have no effect on fitness and hence have their fate dictated by chance alone. The rise of neutralism was supported on two platforms. First, the arrival of protein electrophoresis data implied that polymorphism at the amino acid level was common. This was not predicted by selectionist population genetics, but was expected under neutral theory. Second, Kimura argued that the rate of protein evolution was such that, if all differences between species were owing to selection, the total amount of selective death would be improbably high[2].

While these findings brought the neutral theory to prominence, it has since been a theory in retreat. Neutrality alone cannot explain the level of observed polymorphisms[7]. The theory predicts that species with large populations should show much higher levels of polymorphism than small populations, but this is not observed[7]. Quite why the polymorphism levels are relatively invariant remains unclear, but it is likely to be owing to selection at linked sites, the effect of which is to reduce variation in the vicinity of a gene under positive selection[8].

Another body of evidence against neutrality comes from examination of rates of protein evolution. According to the neutral theory, the number of mutations that become fixed within a population should be Poisson distributed with a mean uT, where T is the number of generations and u is the mutation rate per sequence per generation. This makes two predictions. First, species with short generation times should have faster evolving proteins than those with long generation times. However, this is typically not so[9] and, if there is a molecular clock defined by rates of protein change, it ticks per unit time, not per generation. Second, being Poisson distributed, the mean and variance in the number of substitutions should be equal. On average, however, this is not observed[9]. For example, for non-synonymous (protein changing) mutations in mammals, Ohta[10] estimated that the ratio of the variance to the mean is greater than five (see also[11]). Recent evidence[12] supports the suggestion that this may be owing to episodic positive selection[13].

Perhaps it was unsurprising that protein evolution is not simply neutral. More surprising, however, were investigations of synonymous codon usage. As synonymous nucleotide changes do not alter the encoded amino acid, neutralists argued that they must be invisible to selection[14,15]. Although selectionists noted that, at least in theory, this need not necessarily be true[16], it was not until the early 1980s that evidence emerged for why selection should act at synonymous sites. Studies of bacteria, plants, yeast, fly and worm have revealed that, especially in highly expressed genes, usage of synonymous codons is biased to maximise the rate of protein synthesis by matching skews in tRNA abundances[17-20].

Synonymous mutations in mammals: the last bastion for the neutral theory?

The above organisms all have large populations, so weakly deleterious mutations can be efficiently acted upon by natural selection (BOX 1). However, when populations are small, as in mammals[21] or in species isolated on islands[22], the same mutations can be 'effectively neutral' (BOX 1). Synonymous sites in mammals have, therefore, long been considered to be neutrally evolving[23].

Support for the notion that synonymous mutations in mammals are different is also based on the finding that the dominant factor dictating codon usage in mammals is the isochore effect[4,23,24]. Isochores are large (>300 kb) domains of relatively homogenous guanine+cytosine (GC) content[25]. For a given gene, by far the strongest predictor of nucleotide content at synonymous sites (FIG. 1A) and codon usage bias (FIG. 1B) is the nucleotide content of the isochore[26], i.e. of the flanking non-coding DNA.

The underlying cause of isochoric structure remains uncertain[26], but recent evidence[27-29] suggests that this too is not simply a neutral process. The best current hypothesis (for alternative, see REFS. 30,31) proposes that there exists a mutation bias in favour of A and T, and a fixation bias such that GC content is increased by biased gene conversion, acting either between sister chromosomes during meiotic recombination[32,33] or mitotically between tandem repeats[34].

Are isochore effects alone adequate to explain synonymous codon usage in mammals? The first pieces of evidence we consider may be considered to be indirect tests in that they look for deviations from neutral expectations, while not necessarily specifying a mechanistic basis for the activity of selection. Following this, we review more recent lines of evidence, what we regard as direct evidence, in which specific mechanistic models are examined.

Indirect evidence for selection at synonymous sites

Comparing base composition between synonymous sites within the same gene

Iida and Akashi[35] hypothesised that, because constitutively expressed exons are translated more frequently than alternative exons, a difference in nucleotide content would indicate selection. Indeed, they found in mammals that both GC3 (GC content at the mostly synonymous codon third sites) and the rate of synonymous evolution are higher in exons that are expressed constitutively (N.B. intragenic heterogeneity in synonymous evolution appears to be common[36]). An alternative to comparing constitutive and alternative exons is to assay codon bias within a gene, in a manner that attempts to correct for potential isochore effects[37-39]. Urrutia and Hurst[38], for example, extended a prior method[37] that measures the expected codon usage for each set of synonymous codons, based on the within-gene usage in all the other synonymous sets with the same level of degeneracy. They found that, while isochoric effects do explain

much of the biased codon usage (as expected), they could not explain all of the skew. After correcting for the relationship between codon bias and gene length, the observed codon usage is not correlated with expression breadth[38] but, consistent with selection, is correlated with expression rate [40].

Comparing base composition at synonymous sites with flanking introns

The observation that GC content at synonymous sites is greater than GC in the flanking introns (FIG. 1A), at least for relatively GC-rich regions, could indicate selection at synonymous sites[27,41], not least because the effect may be most pronounced in highly expressed genes[42]. However, this difference can at least in part result from the insertion of AT-rich transposable elements (TEs) into introns within GC-rich isochores[43]. Although reduced, the difference still remains after masking TEs[43]. This residual may be due to the presence of old elements[43], which would be hidden because TEs can only be identified when they have diverged <40% from their progenitor sequence. Nonetheless, this is unlikely to be a complete explanation, as masking elements up to 20% diverged gives almost identical figures[44].

Comparing evolutionary rates at synonymous sites with pseudogenes

If synonymous sites are neutral, they should evolve at the same rate as other putatively neutral sequences. The earliest such tests found that the rate of nucleotide substitutions at synonymous sites is 70% that in pseudogenes[45,46]. Unfortunately, however, such analyses suffer from at least two confounding factors that render interpretation difficult. First, only transcribed genes will experience biases associated with transcriptional-coupled mutation and repair[47,48]. Second, substitution rates vary around the genome (e.g. REF. 49-51), such that related pseudogenes in different locations also evolve at different rates[52,53]. It remains unclear whether either of these factors fully account for the difference[46].

Comparing evolutionary rates at synonymous sites with flanking introns

Performing within-gene analyses[35], such as comparing substitution rates at synonymous sites (K_s) with flanking introns (K_i), avoids the problems of the regional variation in substitution rates and transcription-associated biases. Not all intronic sequence evolves neutrally, however. Both first introns and sequence near intron-exon junctions are conserved by selection[54-56]. While these are relatively easy to exclude, it is hard to define *a priori* those functional regions towards the interior of introns. Consequently, comparing intron evolution with flanking synonymous sites may not prove definitive. Moreover, the hypermutability of CpGs and their differing densities in introns and exons[56] renders comparisons even more problematic. Attempts to exclude the effects of CpGs come to different conclusions[57,58], which may be related to

difficulties in identifying CpG sites[55]. At least in the mouse-rat comparison, differences in rate estimation methods and the difficulty of intron alignment can affect conclusions[59]. Given these difficulties, perhaps then it is not surprising that every possible result has been obtained. Some claim $K_i<K_s$[57,60], others that $K_i=K_s$[41,56,61] and others still that $K_i>K_s$[58,59]. Some suggest that K_s is so much lower than K_i that 40% of synonymous mutations have been opposed by selection[58].

Altogether, these analyses suggest that evolutionary rates alone do not tell us the whole story. Closer analysis is more informative. Notably, even if the overall rates are similar[41,56,61], the patterns of nucleotide substitution at synonymous sites and in introns are quite different[55,56]. For example, C residues are both more common at 4-fold degenerate (synonymous) sites than in introns, and also relatively less likely to be associated with a substitution, after controlling for relative abundance[56] (see also[55]). Further, a claimed reduced rate of synonymous evolution ($K_i>K_s$) is most pronounced on the X-chromosome[62], on which purifying selection is more efficient. The unusually low rate of synonymous evolution in imprinted genes[4] is also then expected as these are also haploid expressed and hence exposed to stronger purifying selection.

Direct tests of specific models of selection

The above evidence, although sometimes contradictory, is nonetheless suggestive of a role for selection. However, an understandable reluctance to accept selection at synonymous sites in mammals must remain until any putative effect is allied with a plausible model.

Maximised translational efficiency

For any given set of synonymous codons, the relevant iso-acceptor tRNAs may not be equally abundant. Consequently, if tRNA abundances are skewed and selection favours rapid translation, there might be a pressure to employ the codon that matches the most abundant tRNA. This model predicts that for any given amino acid there is a 'best' (optimal) codon, defined by the skew in tRNA usage, hence also there must exist a set of codons that should be preferred if translation rate is to be maximised. Co-evolution between biased codon usage and skewed tRNA abundance is possible, leading to a positive feedback loop that exaggerates codon bias and corresponding tRNA skews[63]. Another prediction is that the bias to favour preferred codons should be most pronounced in highly expressed genes and that experimentally adjusted codon usage should affect expression rates. As noted above, these patterns are seen in many organisms[17-20] and, consequently, this translational selection model has been the dominant model for the non-random usage of synonymous codons.

Some data supports a weak relationship between gene expression and codon usage in mammals[40,64,65]. The observed difference in GC content between alternative and constitutive exons[35] was, for example, suggested to support translational selection. However, that certain classes of alternatively spliced exons have low flanking intronic evolution[66] suggests that differences between constitutive and alternative exons may also reflect variation in the density and composition of splicing control elements (see below).

As mentioned above, highly expressed genes exhibit the strongest codon bias[40]. However, correlating bias and expression fails to directly associate codon usage with tRNA abundance (which is reliably assayed by the copy number of tRNA genes[19]). Results of such analyses are contradictory.

Kanaya et al.[67] did not find evidence for skews in putative tRNA genes, while Lander et al.[68] found "only a very rough correlation of human tRNA gene number with either amino-acid frequency or codon bias". Duret[19] interpreted there results as no detectable relationship. Similarly, dos Reis et al.[69] developed a measure of translational selection, *S*, which is the extent to which tRNA copy number and codon usage are co-adapted across genomes. They found that organisms in which selectively driven codon usage bias has previously been described (e.g. *E.coli*, *S.cerevisiae* and *C.elegans*) have high *S*-values ($S>0.45$), whereas humans possessed low values ($S=0.03$), suggesting that selection does not maximise translational efficiency in mammals.

Conversely, two recent studies have found a correlation between tRNA skews and codon usage in humans. Comeron[64], using the data from Lander et al.[68], reports that tRNA copy number matched his proposed set of preferred codons for 14 out of 17 amino acids. Likewise, Lavner et al.[65] show that iso-accepting tRNA numbers positively correlate with expression-weighted frequencies of both amino acids and codons.

Does this mean that adjusting codon usage can modify the rate of translation in mammals, as it does in, for example, *Drosophila*?[70] Numerous studies have demonstrated that modified codon choice can affect net expression levels. Early attempts to express jellyfish green fluorescent protein in human cell lines, for example, were more successful after codon usage was adjusted[71,72] (see also e.g. human/yeast[73]). However, even if in principle translational efficiency can be experimentally maximised by adjusting numerous sites within a gene, it is inappropriate to extrapolate this to supposing that a single synonymous mutation must be under selection, as any given single mutation is unlikely to have a substantial effect on translation rates. Moreover, these experimental results do not always directly demonstrate that it is translation rate that modulates any effect. The transcript must, for example, be efficiently transcribed, have the introns successfully removed and the resulting mRNA must be stable enough to be exported and successfully dock with a ribosome for translation. All of these stages might be sensitive to codon choice. As regards the first possibility, however, support for

a relationship between transcript levels and GC content at silent sites is currently weak[74] and contentious[75]. Evidence for an involvement in mRNA stability and splicing is stronger.

Optimised mRNA stability

If a stable mRNA secondary structure confers resistance to premature degradation, selection might oppose synonymous mutations that disrupt base-pairing[76]. Under this hypothesis, a transcript folds into the optimal conformation given the available sequence, which will for the most part be dictated by protein-coding requirements (note that highly conserved stem-loop sub-structures, e.g. as seen in tRNAs, are unlikely in mammalian mRNAs,[77]). Several cases have highlighted the significance of synonymous mutations affecting mRNA secondary structure[78-80], in some instances associated with disease[79,80] (TABLE 1). Moreover, this model would be consistent with clustering of substitutions within genes[81].

Determining whether synonymous mutations might generally affect fitness, mediated by effects on mRNA folding, is difficult because structures cannot be observed directly. Some studies have, however, investigated the importance of synonymous sites on computationally-predicted mRNA structure and stability in a variety of organisms (e.g. REFS. 82,83). As even *in vitro* foldings may not reflect those formed *in vivo*[84], it is likely that structures predicted *in silico* feature an even larger error component[76]. Nonetheless, recent *in silico* tests in mouse suggest that selection does act at synonymous sites[76]. One particularly intriguing result is that, as previously described in histone genes[85], there is a skew towards G at the first two sites within codons. This can therefore potentially explain the C preference at 4-fold sites[56], as strong G:C pairs create stable mRNAs. Consistent with this are the findings that the stability of wild-type mRNAs relative to artificial transcripts is highest when there is a strong third site skew towards C and mRNAs are less stable when Gs and Cs were interchanged[76]. Moreover, had the synonymous mutations observed in the mouse lineage occurred elsewhere, transcripts would have been less stable[76]. RNA stability therefore provides a possible explanation for C being in excess at third sites.

Transcript stability can also arise from preferring/avoiding particular sequence motifs. Notably, introducing synonymous substitutions that increase C|G dinucleotide content (where | is the codon boundary) decreases the rate of degradation, whereas increasing A|U enhances decay[86]. This may explain avoidance of the AU dinucleotide[38,86], which is cleaved by proteins recognising AU-rich elements[87]. This provides another potential explanation for the C preference at third sites.

Efficient splicing control

That synonymous mutations can be under selection because they upset intron

removal has attracted the most evidence of late. There are abundant examples of synonymous mutations that cause disease by disrupting the splicing process[88,89] (TABLE 1). Nonetheless, such disease-associated mutations are probably much rarer than non-synonymous changes associated with disease, suggesting that only a small fraction of synonymous mutations might have a major effect on splicing. Disease-associated synonymous mutations may create new 'cryptic' splice sites[90] or affect splicing control elements, such as exonic splicing enhancers (ESEs[91]) and silencers (ESSs[92]). Splicing modulators are oligomeric motifs that recruit spliceosomal proteins to facilitate splice site recognition[92]. These tend to be purine-rich[93] and so are unlikely to explain the C excess or its potential association with translation[64] or mRNA stability[76].

Importantly, exonic splicing modulators tend to reside near intron-exon junctions. Much recent evidence has documented the aspects in which the ends of exons are unusual. For example, the codon GAA is common in ESEs and is increasingly preferred over its synonym GAG as one approaches the intron-exon junction[94] (FIG. 2). However, a preference for ESEs, while a robust model, may not explain all the observed gradients in nucleotide content across exons[95]. Alternatively, such biases might reflect an avoidance of codons containing potentially cryptic splice sites[90], those dinucleotides that could be inappropriately identified as intronic ends. If this pressure exists, however, it appears to be much weaker than a preference for ESEs[95].

Consistent with gradients of biased codon choice, some genes exhibit a marked reduction in the rate of synonymous evolution in regions containing an ESE (BRCA1[96-98], FIG. 3; CFTR[99]). More generally, SNP density decreases as one approaches the ends of exons[54], which could be explained by increasing ESE density[100]. Moreover, consistent with purifying selection on ESEs, synonymous SNP frequency is lower in ESEs hexamers than SNP frequency in hexamers from non-exonic sequence[101]. Similarly, synonymous evolution in putative ESEs is slower than in non-ESE sequence, which would largely explains the reduced synonymous substitution rate near exon ends (the authors, unpublished). Selection on exonic splicing modulators may even be more important than the encoded protein[89]. Consistent with this idea, splicing can also affect non-synonymous evolution[102] and amino acid usage (the authors, unpublished).

Implications

The above evidence supports the view that selection acts on some synonymous mutations in mammalian exons. This has important implications. First and foremost, given the evidence for the involvement of synonymous sites in disease, especially mediated by splicing defects (TABLE 1), the assumption of a lack of phenotype of synonymous mutations, like the assumption of neutrality, can no longer be sustained.

Instead of the neutral model, we should then instead be considering synonymous mutations in the framework of the nearly-neutral model (BOX 1). In retrospect, the assumption that synonymous mutations must be neutral because they do not affect protein sequence[14,15] probably reflects the earlier incomplete understanding of the pathway from gene to protein. Indeed, we may still be missing important constraints. It is quite possible, for example, that microRNAs that bind to sense mRNA as a mode of gene regulation might impose constraint on sites in the mRNA to ensure efficient pairing. Further, synonymous mutations can affect protein folding. In *E.coli*, for example, the use of rare codons can induce translational pauses[103] that allow a newly synthesised polypeptide strand enough time to fold into the correct secondary structure[104,105]. Suggestively, stretches of rare codons correspond to turns, loops and links between protein domains[106,107]. Preventing cotranslational misfolding may be even more important in eukaryotes[108] and may explain the preference for GAT over GAC at the N-termini of alpha helices in humans[107]. We also do not yet fully understand why genes expressed uniquely in a given tissue have a GC content that is prototypical for genes expressed in that tissue[31] (N.B. claims that the GC content of tissue-specific genes was independent of isochore effects[109] are not robust, Semon and Duret, unpublished).

Detecting positive selection

One leading use for K_s is as a background evolutionary rate, used to detect positive (Darwinian) selection[110,111]. If selection is favouring adaptive non-synonymous changes, the protein should have a higher rate of evolution than expected under neutral evolution. To this end, the number of non-synonymous substitutions per non-synonymous site (K_a) is compared with K_s. If $K_a > K_s$ then positive selection is inferred, i.e. $K_a/K_s > 1$.

A very low K_s owing to purifying selection on synonymous sites could, in principle, also give rise to $K_a/K_s > 1$ (REF.36,96). This possibility is usually not even considered. However, a few examples have recently been given for intragenic dips in synonymous evolution, most likely associated with splicing regulation[96-99] (FIG. 3). Are these simply oddities or is it the case that an intragenic $K_a/K_s > 1$ often reflects low K_s rather than high K_a? To assess this we examined long (>3000 nucleotides) mouse-rat orthologues and constructed sliding window plots across alignments to search for $K_a/K_s > 1$ peaks. Such peaks are relatively rare, occurring in only 15 of 143 genes. Of the 15, only 11 could be best interpreted as peaks owing to very high K_a with normal K_s or vice versa. The striking conclusion is that 6 could be classified as K_a peaks and 5 as K_s dips (L. D. Hurst, unpublished). This strongly suggests that the K_a/K_s ratio, applied within genes, is not a safe way to identify positive selection, unless one can discount purifying selection on synonymous sites. In principle this might be achieved by

examination of the synonymous rate of evolution in the region with a high Ka/Ks peak so as to ascertain whether the synonymous rate of evolution is unusually low (see also[36]).

Underestimating the mutation rate

If synonymous evolution is not neutral and the synonymous substitution rate (K_s) is used as a measure of the mutation rate, by how much might we be underestimating the true mutation rate? Is it possible to quantify non-neutral effects and hence still use K_s after adjusting for the contribution of selection?

Lu and Wu[62] estimated the proportion of synonymous mutations that are deleterious by comparing rates of evolution between introns and synonymous sites on the X chromosome and the autosome. Remarkably, they estimated that >90% of synonymous mutations are under weak selection. For the most part, however, the selection is so weak that it has a negligible effect on substitution rates. Whether this quantitatively agrees with the 30% lower divergence at synonymous sites compared with pseudogenes[46] is unclear.

An alternative approach is to examine each model individually. However, if one were to quantify the reduction in K_s associated with each model, the relative contributions of each need not be additive. In fly and yeast, there are trade-offs between codon bias for translational efficiency and mRNA secondary structure requirements[112,113]. This caveat aside, given the proportion of exons that specify putative splicing enhancers and the extent to which their rate of evolution is slower than non-ESE sequence, the mutation rate seems to have been underestimated by no more than about 10% (the authors, unpublished), although in one well characterised instance[99], about 30% of synonymous mutations in a given exon are associated with mis-splicing.. A similar quantitative assessment has yet to be performed as regards other modes of selection, although the effects are probably weak. As regards selection at synonymous sites for mRNA stability, only a minority of genes show strong evidence of selection[76] and it probably affects only specific sites. Likewise, codon bias for translational efficiency in mammals, if present, is only detectable in the most highly expressed genes[40]. This suggests that mutation rate estimates are unlikely to increase by more than around 50%. Ambiguity concerning the number of generations that separate taxa, owing to uncertainty in generation times and time since common ancestry, would potentially force adjustments of a much higher order. For example, the time since mouse and rat shared a common ancestor may be anywhere between 5 and 42 million years (for citations see REF. 114).

In short, it is unlikely that the assumption of neutrality of synonymous mutations has grossly misled us in estimates of the genomic mutation rate. This conclusion comes, however, with a strong proviso. Above we asked about selection that

might be peculiar to synonymous mutations. However, aside from the presence of functional residues, there may be reasons to suppose that substitution rates at all silent sites (intronic, intergenic and synonymous) may be misleading. Notably, biased gene conversion will affect substitution rates of all forms of silent DNA[115]. As this process accelerates the fixation of AT→GC mutations and diminishes the rate of fixation of GC→AT mutations, regardless of their coding status, the net rate of evolution will not be equal to the mutation rate, even if the mutations would otherwise be neutral. If the effect is profound, then mutations rates cannot safely be extracted from any sequence comparison.

Optimising transgene expression

Understanding the mode of action of selection on synonymous mutations should enable us to improve transgenes without altering the encoded protein. Although transgene expression is often more efficient when constructs retain the first intron (as these contain regulatory elements), the other introns tend to be dispensable (for citations see REF. 56). In principle, as codon choice near intron-exon junctions is biased to enable efficient splicing[94,95], synonymous sites near junctions could be modified with potentially beneficial effects for transgenes lacking non-first introns. As exonic splice enhancers tend to be A-rich and codon third sites may be C-rich for mRNA stability[76], swapping A for C at synonymous sites might well decrease transcript decay rates. Moreover, a high GC content may also be compatible with the proposed set of preferred codons[64] and will minimise deleterious AU usage[86]. We can foresee that this simple procedure for transgene optimisation could incorporated into a sophisticated *in silico* tool.

Conclusions

The notion that synonymous mutations must be neutral, as they have no effect on the encoded protein, may at first seem both seductive and intuitive. However, the newfound knowledge of what really determines the fate of synonymous mutations in mammals has brought to our attention the unexpected strength of natural selection and a plethora of previously unrecognised selective forces. Acknowledging these will be important for understanding and potentially combating genetic disease. Importantly, understanding how synonymous codon choice makes for efficient expression of a gene will aid in the engineering of better transgenes.

References

1. Kreitman, M. The neutral theory is dead - long live the neutral theory. *Bioessays* **18**, 678-683 (1996).

2. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624-626 (1968).

3. Wolfe, K.H., Sharp, P.M. & Li, W.H. Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283-285 (1989).

4. Smith, N.G.C. & Hurst, L.D. The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? *Genetics* **152**, 661-673 (1999).

5. Eyre-Walker, A. & Keightley, P.D. High genomic deleterious mutation rates in hominids. *Nature* **397**, 344-347 (1999).

6. Keightley, P.D. & Eyre-Walker, A. Deleterious mutations and the evolution of sex. *Science* **290**, 331-333 (2000).

7. Lewontin, R.C. *The Genetic Basis of Evolutionary Change*, (Columbia University Press, New York, 1974).

8. Gillespie, J.H. Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics* **155**, 909-919 (2000).

9. Ohta, T. & Gillespie, J.H. Development of neutral and nearly neutral theories. *Theor. Pop. Biol.* **49**, 128-142 (1996).

10. Ohta, T. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* **40**, 56-63 (1995).

11. Nielsen, R. Robustness of the estimator of the index of dispersion for DNA sequences. *Mol. Phyl. Evol.* 7, 346-351 (1997).

12. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F.J. Erratic overdispersion of three molecular clocks: GPDH, SOD, and XDH. *Proc. Natl Acad. Sci. USA* **98**, 11405-11410 (2001).

13. Gillespie, J.H. *The Causes of Molecular Evolution*, (Oxford University Press, Oxford, 1991).

14. King, J.L. & Jukes, T.H. Non-Darwinian evolution. *Science* **164**, 788-798 (1969).

15. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275-276 (1977).

16. Clarke, B. Darwinian evolution of proteins. *Science* **168**, 1009-1011 (1970).

17. Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 13-34 (1985).

18. Akashi, H. & Eyre-Walker, A. Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8**, 688-693 (1998).

19. Duret, L. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**, 640-649 (2002).

20. Wright, S.I., Yau, C.B., Looseley, M. & Meyers, B.C. Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. *Mol. Biol. Evol.* **21**, 1719-1726 (2004).

21. Keightley, P.D., Lercher, M.J. & Eyre-Walker, A. Evidence for Widespread Degradation of Gene Control Regions in Hominid Genomes. *PLoS Biol.* **3**, e42 (2005).

22. Johnson, K.P. & Seger, J. Elevated rates of nonsynonymous substitution in island birds. *Mol. Biol. Evol.* **18**, 874-881 (2001).

23. Sharp, P.M., Averof, M., Lloyd, A.T., Matassi, G. & Peden, J.F. DNA-sequence evolution: the sounds of silence. *Phil. Trans. R. Soc. Lond. B* **349**, 241-247 (1995).

24. Eyre-Walker, A. An analysis of codon usage in mammals: selection or mutation bias? *J. Mol. Evol.* **33**, 442-449 (1991).

25. Bernardi, G. et al. The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953-958 (1985).

26. Eyre-Walker, A. & Hurst, L.D. The evolution of isochores. *Nat. Rev. Genet.* **2**, 549-555 (2001).

27. Eyre-Walker, A. Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics* **152**, 675-683 (1999).

28. Lercher, M.J., Smith, N.G.C., Eyre-Walker, A. & Hurst, L.D. The evolution of isochores: evidence from SNP frequency distributions. *Genetics* **162**, 1805-1810 (2002).

29. Duret, L., Semon, M., Piganeau, G., Mouchiroud, D. & Galtier, N. Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162**, 1837-1847 (2002).

30. Vinogradov, A.E. Bendable genes of warm-blooded vertebrates. *Mol. Biol. Evol.* **18**, 2195-2200 (2001).

31. Vinogradov, A.E. Isochores and tissue-specificity. *Nucleic Acids Res.* **31**, 5212-5220 (2003).

32. Galtier, N., Piganeau, G., Mouchiroud, D. & Duret, L. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159**, 907-911 (2001).

33. Meunier, J. & Duret, L. Recombination Drives the Evolution of GC-Content in the Human Genome. *Mol. Biol. Evol.* **21**, 984-990 (2004).

34. Galtier, N. Gene conversion drives GC content evolution in mammalian histones. *Trends Genet.* **19**, 65-68 (2003).

35. Iida, K. & Akashi, H. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene* **261**, 93-105 (2000).

36. Pond, S.K. & Muse, S.V. Site-to-Site Variation of Synonymous Substitution Rates. *Mol. Biol. Evol.*, msi232 (2005).

37. Karlin, S. & Mrazek, J. What drives codon choices in human genes? *J. Mol. Biol.* **262**, 459-472 (1996).

38. Urrutia, A.O. & Hurst, L.D. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* **159**, 1191-1199 (2001).

39. Novembre, J.A. Accounting for background nucleotide composition when measuring codon usage bias. *Mol. Biol. Evol.* **19**, 1390-1394 (2002).

40. Urrutia, A.O. & Hurst, L.D. The signature of selection mediated by expression on human genes. *Genome Res.* **13**, 2260-2264 (2003).

41. Hughes, A.L. & Yeager, M. Comparative evolutionary rates of introns and exons in murine rodents. *J. Mol. Evol.* **45**, 125-130 (1997).

42. DeBry, R.W. & Marzluff, W.F. Selection on silent sites in the rodent H3 histone gene family. *Genetics* **138**, 191-202 (1994).

43. Duret, L. & Hurst, L.D. The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. *Mol. Biol. Evol.* **18**, 757-762 (2001).

44. Vinogradov, A.E. Within-intron correlation with base composition of adjacent exons in different genomes. *Gene* **276**, 143-151 (2001).

45. Miyata, T. & Hayashida, H. Extraordinarily high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. *Proc. Natl Acad. Sci. USA* **78**, 5739-5743 (1981).

46. Bustamante, C.D., Nielsen, R. & Hartl, D.L. A maximum likelihood method for analyzing pseudogene evolution: Implications for silent site evolution in humans and rodents. *Mol. Biol. Evol.* **19**, 110-117 (2002).

47. Green, P. et al. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33**, 514-517 (2003).

48. Majewski, J. Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am. J. Hum. Genet.* **73**, 688-692 (2003).

49. Matassi, G., Sharp, P.M. & Gautier, C. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**, 786-791 (1999).

50. Lercher, M.J., Chamary, J.V. & Hurst, L.D. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res.* **14**, 1002-1013 (2004).

51. Gaffney, D.J. & Keightley, P.D. The scale of mutational variation in the murid genome. *Genome Res.* (2005).

52. Casane, D., Boissinot, S., Chang, B.H.J., Shimmin, L.C. & Li, W.H. Mutation pattern variation among regions of the primate genome. *J. Mol. Evol.* **45**, 216-226 (1997).

53. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304 (2000).

54. Majewski, J. & Ott, J. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12**, 1827-1836 (2002).

55. Keightley, P.D. & Gaffney, D.J. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl Acad. Sci. USA* **100**, 13402-13406 (2003).

56. Chamary, J.V. & Hurst, L.D. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: Evidence for selectively driven codon usage. *Mol. Biol. Evol.* **21**, 1014-1023 (2004).

57. Subramanian, S. & Kumar, S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* **13**, 838-844 (2003).

58. Hellmann, I. et al. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13**, 831-837 (2003).

59. Smith, N.G.C. & Hurst, L.D. Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. *J. Mol. Evol.* **47**, 493-500 (1998).

60. Chen, F.C. & Li, W.H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444-456 (2001).

61. Chen, F.C., Vallender, E.J., Wang, H., Tzeng, C.S. & Li, W.H. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J. Hered.* **92**, 481-489 (2001).

62. Lu, J. & Wu, C.I. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proc. Natl Acad. Sci. USA* **102**, 4063-4067 (2005).

63. Bulmer, M. Coevolution of codon usage and transfer RNA abundance. *Nature* **325**, 728-730 (1987).

64. Comeron, J.M. Selective and mutational patterns associated with gene expression in humans: Influences on synonymous composition and intron presence. *Genetics* **167**, 1293-1304 (2004).

65. Lavner, Y. & Kotlar, D. Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene* **345**, 127-138 (2005).

66. Kaufmann, D., Kenner, O., Nurnberg, P., Vogel, W. & Bartelt, B. In NF1, CFTR, PER3, CARS and SYT7, alternatively included exons show higher conservation of surrounding intron sequences than constitutive exons. *Eur. J. Hum. Genet.* **12**, 139-149 (2004).

67. Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. & Ikemura, T. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.* **53**, 290-298 (2001).

68. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

69. dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036-5044 (2004).

70. Carlini, D.B. & Stephan, W. In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. *Genetics* **163**, 239-243 (2003).

71. Levy, J.P., Muldoon, R.R., Zolotukhin, S. & Link, C.J., Jr. Retroviral transfer and expression of a humanized, red-shifted green fluorescent protein gene into human tumor cells. *Nat. Biotech.* **14**, 610-614 (1996).

72. Zolotukhin, S., Potter, M., Hauswirth, W.W., Guy, J. & Muzyczka, N. A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. *J. Virol.* **70**, 4646-4654 (1996).

73. Kim, C.H., Oh, Y. & Lee, T.H. Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells. *Gene* **199**, 293-301 (1997).

74. Lercher, M.J., Urrutia, A.O., Pavlicek, A. & Hurst, L.D. A unification of mosaic structures in the human genome. *Hum. Mol. Genet.* **12**, 2411-2415 (2003).

75. Semon, M., Mouchiroud, D. & Duret, L. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Hum. Mol. Genet.* **14**, 421-427 (2005).

76. Chamary, J.V. & Hurst, L.D. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* **6**, R75 (2005).

77. Buratti, E. & Baralle, F.E. Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol. Cell. Biol.* **24**, 10505-10514 (2004).

78. Shen, L.X., Basilion, J.P. & Stanton, V.P., Jr. Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc. Natl Acad. Sci. USA* **96**, 7871-7876 (1999).

79. Duan, J. et al. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* **12**, 205-216 (2003).

80. Capon, F. et al. A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Hum. Mol. Genet.* **13**, 2361-2368 (2004).

81. Smith, N.G. & Hurst, L.D. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153**, 1395-1402 (1999).

82. Seffens, W. & Digby, D. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.* **27**, 1578-1584 (1999).

83. Cohen, B. & Skiena, S. Natural selection and algorithmic design of mRNA. *J. Comp. Biol.* **10**, 419-432 (2003).

84. Schroeder, R., Barta, A. & Semrad, K. Strategies for RNA folding and assembly. *Nat. Rev. Mol. Cell Biol.* **5**, 908-919 (2004).

85. Huynen, M.A., Konings, D.A. & Hogeweg, P. Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure. *J. Mol. Evol.* **34**, 280-291 (1992).

86. Duan, J. & Antezana, M.A. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. *J. Mol. Evol.* **57**, 694-701 (2003).

87. Beutler, E., Gelbart, T., Han, J.H., Koziol, J.A. & Beutler, B. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proc. Natl Acad. Sci. USA* **86**, 192-196 (1989).

88. Cartegni, L., Chew, S.L. & Krainer, A.R. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nat. Rev. Genet.* **3**, 285-298 (2002).

89. Pagani, F. & Baralle, F.E. Genomic variants in exons and introns: identifying the splicing spoilers. *Nat. Rev. Genet.* **5**, 389-396 (2004).

90. Eskesen, S.T., Eskesen, F.N. & Ruvinsky, A. Natural selection affects frequencies of AG and GT dinucleotides at the 5' and 3' ends of exons. *Genetics* **167**, 543-550 (2004).

91. Fairbrother, W.G., Yeh, R.F., Sharp, P.A. & Burge, C.B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007-1013 (2002).

92. Wang, Z. et al. Systematic identification and analysis of exonic splicing silencers. *Cell* **119**, 831-845 (2004).

93. Blencowe, B.J. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* **25**, 106-110 (2000).

94. Willie, E. & Majewski, J. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* **20**, 534-538 (2004).

95. Chamary, J.V. & Hurst, L.D. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? *Trends Genet.* **21**, 256-259 (2005).

96. Hurst, L.D. & Pal, C. Evidence for purifying selection acting on silent sites in BRCA1. *Trends Genet.* **17**, 62-65 (2001).

97. Liu, H.X., Cartegni, L., Zhang, M.Q. & Krainer, A.R. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat. Genet.* **27**, 55-58 (2001).

98. Orban, T.I. & Olah, E. Purifying selection on silent sites - a constraint from splicing regulation? *Trends Genet.* **17**, 252-253 (2001).

99. Pagani, F., Raponi, M. & Baralle, F.E. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl Acad. Sci. USA* **102**, 6368-6372 (2005).

100. Fairbrother, W.G., Holste, D., Burge, C.B. & Sharp, P.A. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* **2**, e268 (2004).

101. Carlini, D.B. & Genut, J.E. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Evol.* **In press**(2005).

102. Cusack, B.P. & Wolfe, K.H. Changes in Alternative Splicing of Human and Mouse Genes are Accompanied by Faster Evolution of Constitutive Exons. *Mol. Biol. Evol.* (2005).

103. Purvis, I.J. et al. The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. *J. Mol. Biol.* **193**, 413-417 (1987).

104. Komar, A.A., Lesnik, T. & Reiss, C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Letters* **462**, 387-391 (1999).

105. Cortazzo, P. et al. Silent mutations affect in vivo protein folding in Escherichia coli. *Biochem. Biophys. Res. Comm.* **293**, 537-541 (2002).

106. Thanaraj, T.A. & Argos, P. Ribosome-mediated translational pause and protein domain organization. *Protein Sci.* **5**, 1594-1612 (1996).

107. Oresic, M. & Shalloway, D. Specific correlations between relative synonymous codon usage and protein secondary structure. *J. Mol. Biol.* **281**, 31-48 (1998).

108. Netzer, W.J. & Hartl, F.U. Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* **388**, 343-349 (1997).

109. Plotkin, J.B., Robins, H. & Levine, A.J. Tissue-specific codon usage and the expression of human genes. *Proc. Natl Acad. Sci. USA* **101**, 12588-12591 (2004).

110. Yang, Z.H. & Bielawski, J.P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496-503 (2000).

111. Hurst, L.D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18**, 486 (2002).

112. Carlini, D.B., Chen, Y. & Stephan, W. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. *Genetics* **159**, 623-633 (2001).

113. Carlini, D.B. Context-Dependent Codon Bias and mRNA Longevity in the Yeast Transcriptome. *Mol. Biol. Evol.* (2005).

114. Adkins, R.M., Gelke, E.L., Rowe, D. & Honeycutt, R.L. Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. *Mol. Biol. Evol.* **18**, 777-791 (2001).

115. Piganeau, G., Mouchiroud, D., Duret, L. & Gautier, C. Expected relationship between the silent substitution rate and the GC content: Implications for the evolution of isochores. *J. Mol. Evol.* **54**, 129-133 (2002).

116. Denecke, J., Kranz, C., Kemming, D., Koch, H.G. & Marquardt, T. An activated 5 ' cryptic splice site in the human ALG3 gene generates a premature termination codon insensitive to nonsense-mediated mRNA decay in a new case of congenital disorder of glycosylation type Id (CDG-Id). *Hum. Mut.* **23**, 477-486 (2004).

117. Montera, M. et al. A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. *J. Med. Genet.* **38**, 863-867 (2001).

118. Aretz, S. et al. Familial adenomatous polyposis: Aberrant splicing due to missense or silent mutations in the APC gene. *Hum. Mut.* **24**, 370-380 (2004).

119. Hellwinkel, O.J.C. et al. A unique exonic splicing mutation in the human androgen receptor gene indicates a physiologic relevance of regular androgen receptor transcript variants. *J. Clin. Endocrinol. Metab.* **86**, 2569-2575 (2001).

120. Teraoka, S.N. et al. Splicing defects in the ataxia-telangiectasia gene, ATM: Underlying mutations and consequences. *Am. J. Hum. Genet.* **64**, 1617-1631 (1999).

121. O'Driscoll, M., Ruiz-Perez, V.L., Woods, C.G., Jeggo, P.A. & Goodship, J.A. A splicing mutation affecting expression of ataxia-telangiectasia and Rad3-related protein (ATR) results in Seckel syndrome. *Nat. Genet.* **33**, 497-501 (2003).

122. Ishibashi, F. et al. Improved superoxide-generating ability by interferon {gamma} due to splicing pattern change of transcripts in neutrophils from patients with a splice site mutation in CYBB gene. *Blood* **98**, 436-441 (2001).

123. Chen, W. et al. Silent nucleotide substitution in the sterol 27-hydroxylase gene (CYP 27) leads to alternative pre-mRNA splicing by activating a cryptic 5 ' splice

site at the mutant codon in cerebrotendinous xanthomatosis patients. *Biochemistry* **37**, 4420-4428 (1998).

124. vanAmstel, J.K.P. et al. Hereditary tyrosinemia type 1: Novel missense, nonsense and splice consensus mutations in the human fumarylacetoacetate hydrolase gene; Variability of the genotype-phenotype relationship. *Hum. Genet.* **97**, 51-59 (1996).

125. Liu, W.G., Qian, C.P. & Francke, U. Silent mutation induces exon skipping of fibrillin-1 gene in Marfan syndrome. *Nat. Genet.* **16**, 328-329 (1997).

126. Flusser, H. et al. Mild glycine encephalopathy (NKH) in a large kindred due to a silent exonic GLDC splice mutation. *Neurology* **64**, 1426-1430 (2005).

127. Harteveld, C.L. et al. An alpha-thalassemia phenotype in a Dutch Hindustani, caused by a new point mutation that creates an alternative splice, donor site in the first Exon of the alpha 2-globin gene. *Hemoglobin* **28**, 255-259 (2004).

128. Akli, S. et al. A G-Mutation to a-Mutation at Position-1 of a 5' Splice Site in a Late Infantile Form of Tay-Sachs Disease. *J. Biol. Chem.* **265**, 7324-7330 (1990).

129. Wicklow, B.A. et al. Severe subacute G(M2) gangliosidosis caused by an apparently silent HEXA mutation (V324V) that results in aberrant splicing and reduced HEXA mmRNA. *Am. J. Med. Genet. Part A* **127A**, 158-166 (2004).

130. Llewellyn, D.H. et al. Acute intermittent porphyria caused by defective splicing of porphobilinogen deaminase RNA: A synonymous codon mutation at - 22 bp from the 5' splice site causes skipping of exon 3. *J. Med. Genet.* **33**, 437-438 (1996).

131. Valentine, C.R. The association of nonsense codons with exon skipping. *Mutat. Res.-Rev. Mutat. Res.* **411**, 87-117 (1998).

132. Jin, Y. et al. Glanzmann thrombasthenia - Cooperation between sequence variants in cis during splice site selection. *J. Clin. Invest.* **98**, 1745-1754 (1996).

133. Xie, J.L., Pabon, D., Jayo, A., Butta, N. & Gonzalez-Manchon, C. Type I Glanzmann thrombasthenia caused by an apparently silent beta 3 mutation that results in aberrant splicing and reduced beta 3 mRNA. *Thromb. Haemost.* **93**, 897-903 (2005).

134. Buchroithner, B. et al. Analysis of the LAMB3 gene in a junctional epidermolysis bullosa patient reveals exonic splicing and allele-specific nonsense-mediated mRNA decay. *Laboratory Investigation* **84**, 1279-1288 (2004).

135. Du, Y.Z., Dickerson, C., Aylsworth, A.S. & Schwartz, C.E. A silent mutation, C924T (G308G), in the L1CAM gene results in X linked hydrocephalus (HSAS). *J. Med. Genet.* **35**, 456-462 (1998).

136. Ries, S. et al. Different missense mutations in histidine-108 of lysosomal acid lipase cause cholesteryl ester storage disease in unrelated compound heterozygous and hemizygous individuals. *Hum. Mut.* **12**, 44-51 (1998).

137. D'Souza, I. et al. Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17†type, by affecting multiple alternative RNA splicing regulatory elements. *Proc. Natl Acad. Sci. USA* **96**, 5598-5603 (1999).

138. Spillantini, M.G. et al. A novel tau mutation (N296N) in familial dementia with swollen achromatic neurons and corticobasal inclusion bodies. *Ann. Neurol.* **48**, 939-943 (2000).

139. Stanford, P.M. et al. Progressive supranuclear palsy pathology caused by a novel silent mutation in exon 10 of the tau gene: Expansion of the disease phenotype caused by tau gene mutations. *Brain* **123**, 880-893 (2000).

140. KohonenCorish, M. et al. RNA-based mutation screening in hereditary nonpolyposis colorectal cancer. *Am. J. Hum. Genet.* **59**, 818-824 (1996).

141. Fahsold, R. et al. Minor lesion mutational spectrum of the entire NF1 gene does not explain its high mutability but points to a functional domain upstream of the CAP-related domain. *Am. J. Hum. Genet.* **66**, 790-818 (2000).

142. Chao, H.K., Hsiao, K.J. & Su, T.S. A silent mutation induces exon skipping in the phenylalanine hydroxylase gene in phenylketonuria. *Hum. Genet.* **108**, 14-19 (2001).

143. Cardozo, A.K., De Meirleir, L., Liebaers, I. & Lissens, W. Analysis of exonic mutations leading to exon skipping in patients with pyruvate dehydrogenase E1 alpha deficiency. *Pediatr. Res.* **48**, 748-753 (2000).

144. Kanno, H. et al. Frame shift mutation, exon skipping, and a two-codon deletion caused by splice site mutations account for pyruvate kinase deficiency. *Blood* **89**, 4213-4218 (1997).

145. Jacobsen, M. et al. A point mutation in PTPRC is associated with the development of multiple sclerosis. *Nat. Genet.* **26**, 495-499 (2000).

146. Imamura, T., Okano, Y., Shintaku, H., Hase, Y. & Isshiki, G. Molecular characterization of 6-pyruvoyl-tetrahydropterin synthase deficiency in Japanese patients. *J. Hum. Genet.* **44**, 163-168 (1999).

147. Fernandez-Cadenas, I. et al. Splicing mosaic of the myophosphorylase gene due to a silent mutation in McArdle disease. *Neurology* **61**, 1432-1434 (2003).

148. Auricchio, A. et al. Double heterozygosity for a RET substitution interfering with splicing and an EDNRB missense mutation in Hirschsprung disease. *Am. J. Hum. Genet.* **64**, 1216-1221 (1999).

149. Lorson, C.L., Hahnen, E., Androphy, E.J. & Wirth, B. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc. Natl Acad. Sci. USA* **96**, 6307-6311 (1999).

150. Mizuguchi, T. et al. Heterozygous TGFBR2 mutations in Marfan syndrome. *Nat. Genet.* **36**, 855-860 (2004).

151. Ferrari, S. et al. Mutations of CD40 gene cause an autosomal recessive form of immunodeficiency with hyper IgM. *Proc. Natl Acad. Sci. USA* **98**, 12614-12619 (2001).

152. Mendez, M. et al. Familial porphyria cutanea tarda: Characterization of seven novel uroporphyrinogen decarboxylase mutations and frequency of common hemochromatosis alleles. *Am. J. Hum. Genet.* **63**, 1363-1375 (1998).

BOX 1. **The neutral theory, the nearly-neutral theory and why mammals might be different.**

The strictly neutral theory considers the fate of mutations that have no effect on fitness. If such mutations occur at a rate μ per haploid genome per generation, then each generation there must be $2N\mu$ new neutral mutations in a diploid population of size N. What is the fate of any such new mutation? Random fluctuations in allele frequency (drift) permit the new mutation to go up or down in frequency. The chance that the new mutation will become fixed in a population is $1/2N$, i.e. the same as pulling one white ball from a collection of $2N$ balls where only one is white. Consequently, the neutral mutation rate (μ) is the rate of fixation, as $2N\mu/2N = \mu$. Hence evolution at neutral sites can be employed to estimate the mutation rate.

What if a mutation has only a small effect on fitness? The successor to strict neutrality, the nearly-neutral theory[9], considers the fate of such mutations. The theory predicts that a mutation will be 'effectively neutral' if its selective disadvantage (s) is small compared to the effective population size, N_e, (more precisely, if $s<<1/2N_e$,[1]). By 'effectively neutral', we mean that the fixation rate is so close to μ as to make no difference. By contrast, if a mutation is 'slightly deleterious', it can be opposed by selection if the fitness effect is larger or the population size smaller (with $s\approx1/2N_e$), while still permitting substitutions at some measurable rate, i.e. a fixation rate less than μ. If the mutation is even more deleterious ($s>>1/2\ N_e$), then the mutation will not reach fixation. Mutations that cause evident disease are just the more extreme examples of those incapable of reaching fixation.

Note that what classifies as a slightly deleterious mutation is dependent on the effective population size. A mutation in a fly could be slightly deleterious ($s\approx1/2\ N_e$), while one of the same fitness in a mammal could be effectively neutral ($s<1/2N_e$). It was hence argued that it would be unlikely for selection to affect synonymous mutations in species with small populations[23], such as mammals, where $N_e<<10^6$ (REF. 21), while still affecting codon usage in bacteria, fly, etc. The nearly-neutral theory correctly predicts lower levels of selective constraint in small populations[6].

Table 1. **Human diseases associated with synonymous mutations (based on REF. 88).**

Gene	Mutation	Exon	Mechanism	Disease	Ref
ALG3	G55G	1	ESE activates upstream cryptic SS?	Congenital disorder of glycosylation type Id	116
APC	R623R	14	ESE disrupted?	Familial adenomatous polyposis (FAP)	117
	H652H	14	ESE disrupted?		118
	R653R	14	ESE disrupted?		118
AR	S888S	8	5' SS created	Androgen insensitivity syndrome (AIS)	119
ATM	S706S	16	5' SS disrupted	Ataxia-telangiectasia (AT)	120
ATM	S1135S	26	5' SS disrupted	Ataxia-telangiectasia (AT)	120
ATR	G677G	9	mRNA structure?	Seckel syndrome	121
CYBB	A84A	3	5' SS disrupted	Chronic granulomatous disease (CGD)	122
CYP27A1	G112G	2	5' SS created	Cerebrotendinous Xanthomatosis	123
FAH	N232N	8	Unknown	Hereditary tyrosinemia type 1 (HT 1)	124
FBN1	I2118I	51	Unknown	Marfan syndrome	125
GLDC	P869P	22	ESE?	Glycine encephalopathy (NKH)	126
HBA2	G22G	1	5' SS created	An α-thalassemia disease	127
HEXA	L187L	5	5' SS disrupted	Tay-Sachs disease	128
HEXA	V324V	8	5' SS created	G_{M2} Gangliosidosis	129
HMBS	R28R	3	ESE disrupted?	Acute intermittent porphyria (AIP)	130
HPRT1	F199F	8	Unknown	Potenially Lesch-Nyhan syndrome	131
ITGB3	T420T	9	mRNA structure?	Glanzmann thrombasthenia	132
ITGB3	G605G	11	5' SS created	Glanzmann thrombasthenia	133
LAMB3	H1003H	20	5' SS created	Junctional epidermolysis bullosa	134
L1CAM	G308G	8	5' SS created	X linked hydrocephalus (HSAS)	135
LIPA	Q277Q	8	Unknown	Cholesteryl ester storage disease	136
MAPT	L284L	10	ESE or ESS disrupted	Frontotemporal dementia with parkinsoniam-chromosome 17 type (FTDP-17)	137
MAPT	N296N	10	ESS disrupted	Familial dementia with swollen achromatic neurons and corticobasal inclusion bodies	138
MAPT	S305S	10	5' SS disrupted	Supranuclear palsy	139
MLH1	S577S	16	Unknown	Hereditary nonpolyposis colorectal cancer	140
NF1	K354K	7	5' SS disrupted	Neurofibromatosis type 1 (NF1)	141
PAH	V399V	11	ESE disrupted?	Phenylketonuria	142
PDHE1	G185G	6	ESE disrupted	Leigh's encephalomyelopathy	143
PKLR	A423A	9	Unknown	Pyruvate Kinase deficiency	144
PTPRC	P48P	4	Unknown	Multiple sclerosis (MS)	145
PTS	E81E	4	5' SS disrupted	PTPS deficiency	146
PYGM	K608K	15	Unknown	McArdle disease	147
RET	I647I	11	ESE?	Hirschsprung disease	148
SMN1	F280F	7	ESE disrupted	Spinal muscular atrophy (SMA)	149
TGFBR2	Q508Q	6	5' SS disrupted	Marfan syndrome	150
TNFRSF5	T136T	5	ESE disrupted	Immunodeficiency with hyper IgM	151
UROD	E314E	9	5' SS disrupted	Familial Porphyria Cutanea Tarda (f-PCT)	152

Figure 1A. **The relationship between GC content in introns and at codon third sites in 1380 human genes (R^2=0.60, P<0.0001).** The line indicates equality.

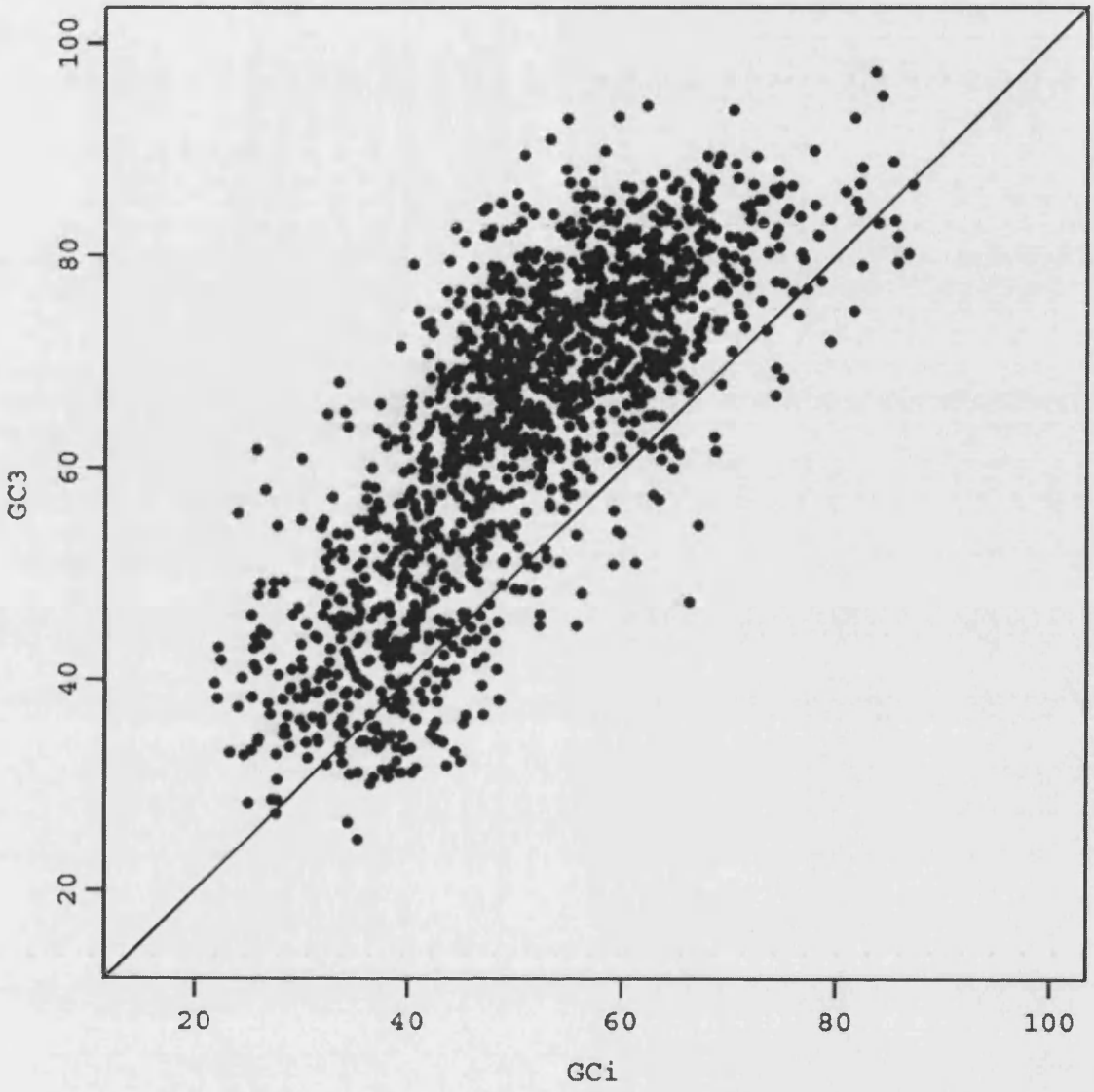


Figure 1B. **The relationship between GC content at codon third sites and codon usage bias (measured by the effective number of codons) for 2030 human genes.** The lower the effective number of codons, the greater the codon usage bias.

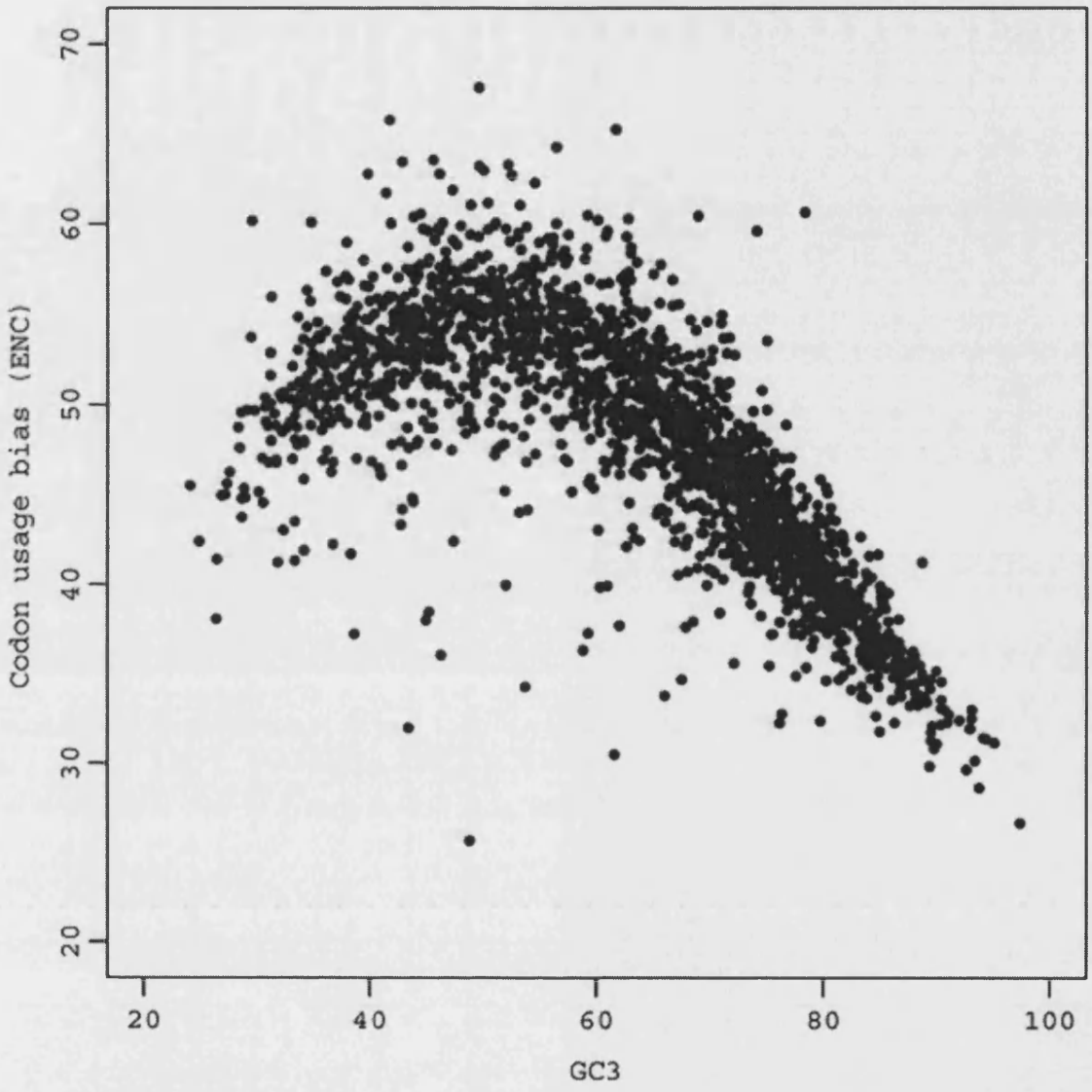


Figure 2. **Proportional usage of GAA versus its synonym GAG, as a function of the distance from the (5' or 3') ends of 14407 human exons (1802 genes, R^2=0.88, P<0.0001)**

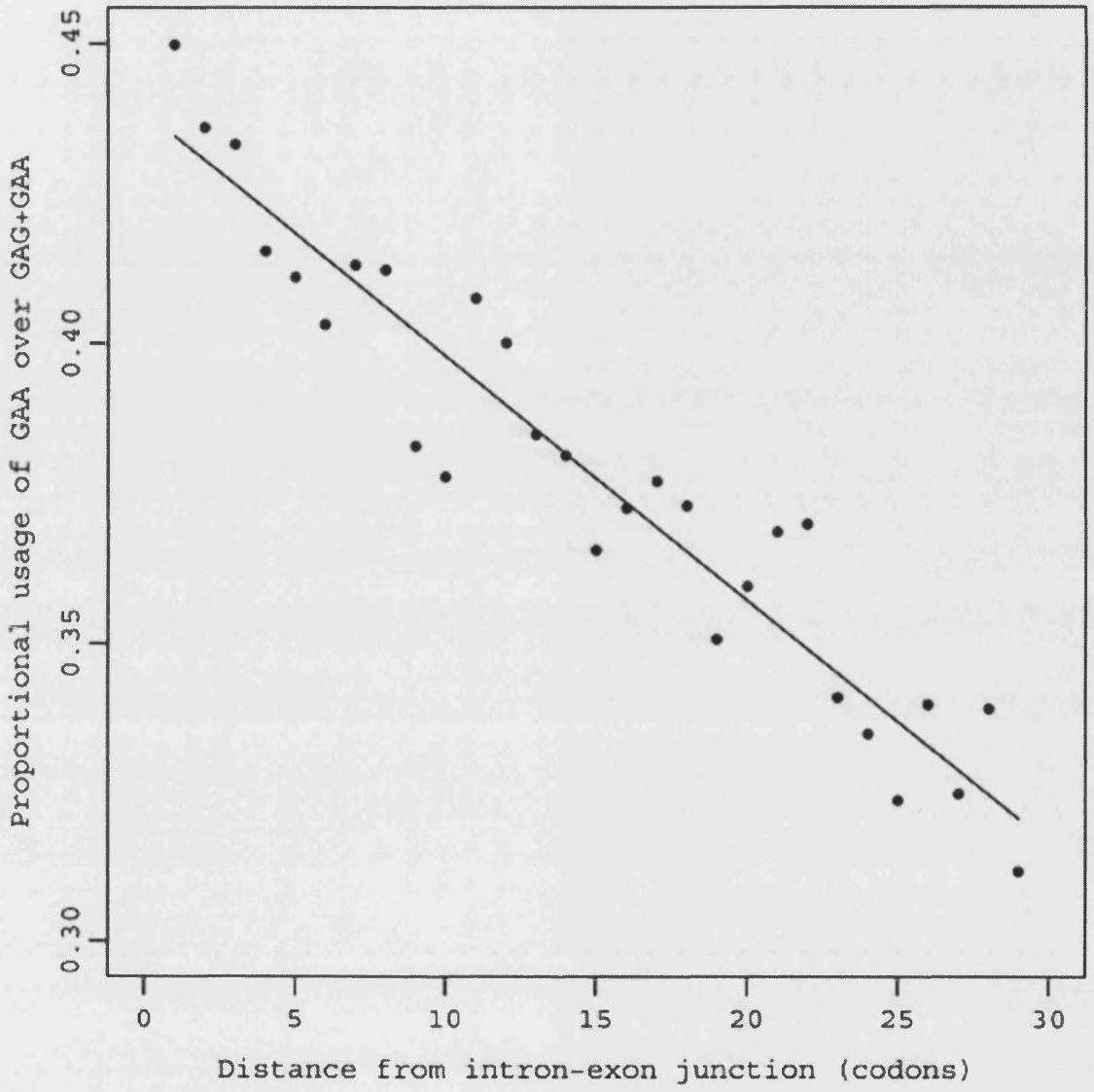
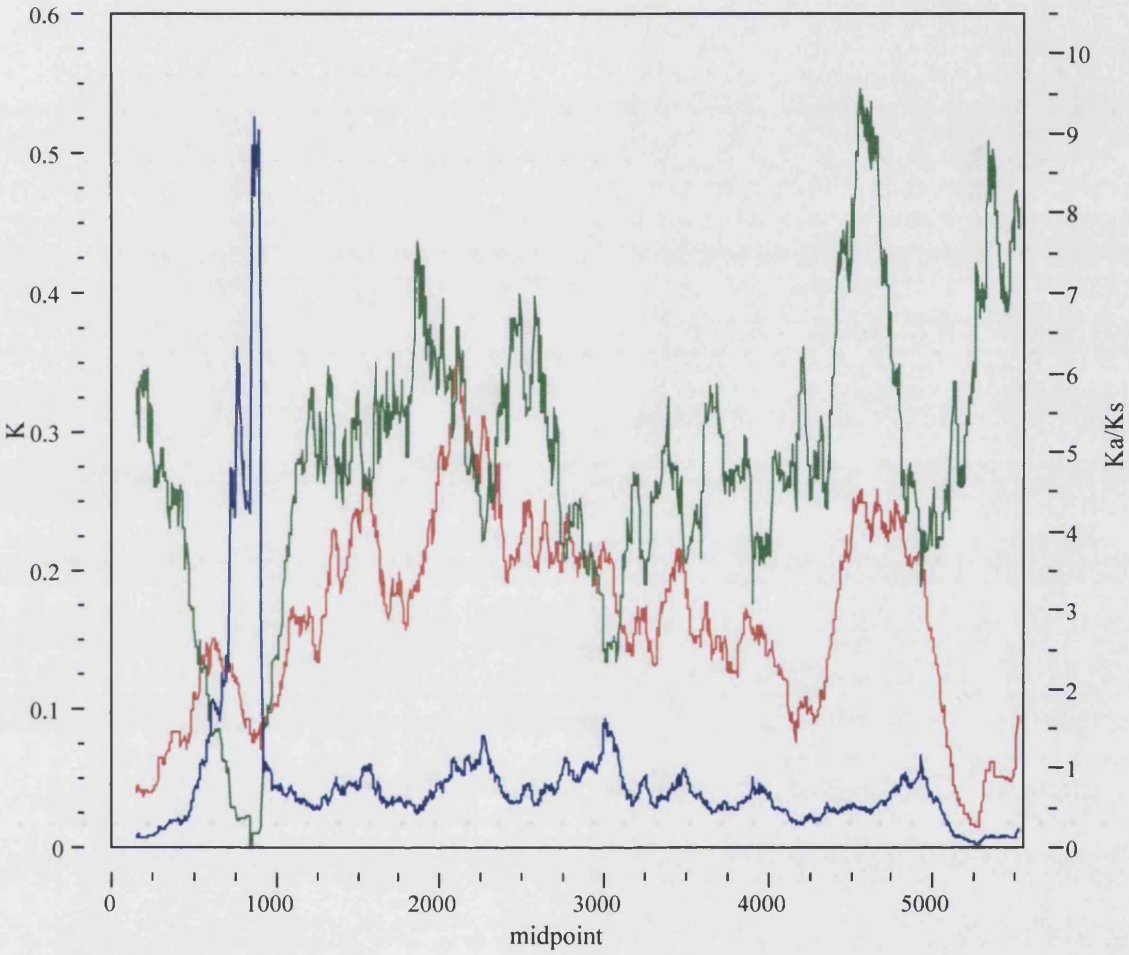


Figure 3. **A sliding window plot of rates of evolution across the BRCA1 gene (compared between human and dog).** The x-axis shows the midpoint of the 306 nucleotide window. The y-axis shows the rate of non-synonymous substitution (K_a, red), the rate of synonymous evolution (K_s, green) and the K_a/K_s ratio (blue). Note that the high K_a/K_s peak near the 5' end of the gene is associated with a striking dip in K_s rather than a peak in K_a.



Chapter 8. **Discussion**

The story so far

"In this new era of molecular biology, it should not be surprising if evolutionary theory is subject to some revision."
(Kimura & Ohta 1973)

"For molecular evolution researchers, the genomics revolution has showered us with raw data [...], enabling us to infer genome-wide evolutionary patterns"
(Wolfe & Li 2003)

Although research in molecular evolution has recently made the transition into the post-genomic era, many of the burning questions remain the same as those asked over 30 years ago. Today, however, large datasets of both coding and non-coding sequence have allowed us to address how evolutionary forces act at the level of whole genomes. Mutation, as the major force generating genotypic and ultimately phenotypic variation, continues to be of particular interest. Recent articles in high-impact journals have demonstrated, for example, the extent to which Darwinian selection drives amino acid and protein evolution (Clark et al. 2003; Bazykin et al. 2004; Nielsen et al. 2005).

Given the arguments of neutral theory (Chapter 1) and the peculiarities of mammalian genomes (Chapter 7), it might seem controversial to suggest that natural selection is powerful enough to influence the fate of mutations that do not alter amino acids. My contribution to a "revision" of evolutionary theory has been to show that a proportion of silent sites are indeed under selection, even in mammals. As this implies that they must have an impact upon fitness, 'silent' evolution should be studied under the framework of nearly-neutral (rather than strictly neutral) theory.

In Part I of this thesis, I reported that rates of evolution in introns and at synonymous sites vary across the mammalian genome (Chapter 2). Importantly, local similarity in evolutionary rates is not an artefact of analysing genes exhibiting "disparity" in the patterns of evolution between the orthologues in each gene pair. Had such regional variation not been observed, one would have had to accept the strictly neutral (null) hypothesis that there is one mutation rate for all genes, and that substitution rates at silent sites provide a measure of the point mutation rate (Kumar & Subramanian 2002). The finding that silent sites in separate autosomal genes evolve at different rates has two interpretations, and these are not mutually exclusive. Firstly, it could be caused by variation in the mutation rate across the genome (Filipski 1988; Ellegren, Smith & Webster 2003). If this is the case, it may be due to recombination-

induced mutagenesis (e.g. Perry & Ashworth 1999; Hellmann et al. 2003). The second explanation is that the level of selective constraint at silent sites differs between genes.

In Part II, I showed that both introns and exonic silent sites are subject to selection (Chapter 3). For introns, the ends are conserved between mouse and rat (presumably to preserve sites required for splice site recognition), while first introns evolve more slowly than the other introns in the same gene (as the former are enriched for elements that control gene regulation). After eliminating selectively constrained intronic sequence, I then performed another test of neutrality at silent sites. Contrary to neutral expectations, I found that introns and synonymous sites differed in their patterns of evolution. The main result of the study came from comparing the frequency of nucleotide substitutions at four-fold synonymous and intronic sites, controlling for the relative abundance of the dinucleotide context in which they occur. I found that A and T are relatively unstable at third sites in exons, but that C was particularly stable. Under the supposition that most of the remaining intronic sequence evolves neutrally, this suggests that codon usage is influenced by selection in murid rodents.

In Part III, I investigated two mechanistic explanations for observed biases in synonymous codon usage. As evidence to support the concept that selection chooses codons that maximise the efficiency of protein synthesis (Akashi & Eyre-Walker 1998) remains controversial in mammals (see Chapter 7 and Duret 2002), I examined some of the alternative models instead.

Notably, the preference for C at third sites (Chapter 3) can be explained by selection to optimise the stability of mRNA secondary structure (Chapter 4). Through various randomisation protocols, I also found that, no matter how one modifies an mRNA sequence, real transcripts tend to be more stable than expected by chance. For example, by simulating evolution and reallocating the substitutions observed in the mouse lineage, I found that had synonymous mutations occurred at different locations along mRNAs, they would on average have generated transcripts with lower stability.

Selection can also act at synonymous sites to maintain efficient splicing. This latter function, explaining observed gradients in codon choice near intron-exon junctions, can be split into two models. First, certain codons might be avoided so that the spliceosome does not inappropriately recognise them as cryptic splice sites (Eskesen, Eskesen & Ruvinsky 2004). Second, exonic splicing enhancers (ESEs) are common near junctions, which might lead to certain codons being preferred (Willie & Majewski 2004). Some effects are expected under both models, so I identified and tested their discriminating predictions (Chapter 5). As I found strong support for the latter model, I then quantified the level of selective pressure on synonymous sites required to preserve ESEs (Chapter 6).

In the penultimate chapter (Chapter 7), I reviewed the evidence and highlighted some of the important implications of the finding that synonymous sites are under natural selection in mammals.

Where next?

Variation in silent site substitution rates among regions of the mammalian genome is a genuine evolutionary pattern (Chapter 2). To date, no study has been able to support Kumar and Subramanian's (2002) assertion that there is a single rate shared by all autosomal sequences. Indeed, a recent use of a non-comparative method casts further doubt on their claim that regional variation can be explained by the analysis of orthologues in which members of a pair are evolving disparately. Arndt et al. (2005) scanned the human genome for repetitive elements, comparing them to the ancestral "master sequences" to examine the mutations that have accumulated. The authors found substantial regional variation after analysing each type of substitution independently (in line with the idea that substitution rates alone do not provide enough information on evolutionary processes, Chapter 3).

For local similarity, can estimating the scale of the effect provide clues as to what causes mutation rates to vary across the genome? Gaffney and Keightley (2005) used the substitution rate in ancestral repeats to assign a value of 1Mb to the size of the "evolutionary rate units" (Matassi, Sharp & Gautier 1999) of local similarity in the murid genome. Like isochores (Eyre-Walker & Hurst 2001), however, these "units" are not discrete blocks. The effect decays with distance, so that local similarity cannot be detected when block size exceeds 15Mb. Contrary to the suggestion that local similarity might be caused by recombination-induced mutation (Chapter 2), the authors do not find a correspondence between substitution and recombination rates. Nonetheless, they suggest that this is still the best mechanistic explanation. The relationship may still exist, but is complicated by the two processes occurring at different physical (megabase versus kilobase) and temporal scales. Moreover, in humans, the majority of recombination is restricted to relatively small "hotspots" (Crawford et al. 2004; McVean et al. 2004) which can shift in position over relatively short periods of time (e.g. since humans and chimps diverged, Ptak et al. 2005).

The megabase scale for local similarity demonstrates that it is a sub-chromosomal rather than inter-chromosomal effect. Navarro and Barton (2003) suggested that intra-chromosomal variation in rates of evolution can be caused by 'chromosomal speciation', whereby rearrangements (e.g. inversions) reduce gene flow and lead to reproductive isolation between individuals. By comparing ~100 human-chimp genes, the authors found that divergence in rearranged (R) regions was higher than in co-linear (C) regions, which might indicate suppressed recombination (i.e.

relaxed selection) in the former. However, R>C evolution was also found in the human-gorilla comparison (Lu, Li & Wu 2003), and a genome-wide human-chimp analysis found no difference between the two region types (Mikkelsen et al. 2005). Moreover, R and C regions do not differ in their rates of human-chimp expression divergence (Zhang, Wang & Podlaha 2004). Hence this model cannot explain regional variation.

Introns can contain elements that may regulate gene expression (Chapter 3). Whether selection is strong enough to conserve them is another matter, however. Keightley et al. (2005) showed that, in the hominid lineage, putative non-coding control regions proximal to the 5' end of genes (including first introns, Chapter 3) are accumulating mildly deleterious mutations. Murids, by contrast, are better able to conserve these functional elements. The most likely explanation, they suggest, is that the strength of purifying selection is weaker in hominids compared to murids. This may be due to an order of magnitude difference in their effective population sizes (approximately 20,000 versus 600,000, Keightley, Lercher & Eyre-Walker 2005).

Functional elements within introns might also be active after splicing. Mattick argues that introns are replete with functional non-coding RNAs (ncRNAs), suggesting that their role in regulating gene expression (Mattick & Makunin 2005) represents a "new genetics" (Mattick 2004). Functional transcription units cover at least 50% of the human genome, of which a third do not encode proteins (Semon & Duret 2004). If this ~15% represents functional ncRNA genes, we may have grossly underestimated the proportion of silent DNA under selection (but see Keightley, Lercher & Eyre-Walker 2005). At present, however, the identification of ncRNAs is still in its infancy (Kapranov et al. 2002; Okazaki et al. 2002; Kampa et al. 2004), as is the development of bioinformatics tools to predict their location (e.g. Washietl, Hofacker & Stadler 2005). Conclusions drawn from indirect tests of neutrality (Chapters 2 and 3) may need to be revised if a large fraction of intronic sequence is under selection, although note that such comparative tests can sometimes be uninformative (Chapter 7).

Evolution at synonymous sites and codon usage bias was extensively covered in the previously (Chapter 7), which described the most recent developments in this area. While I have suggested that non-neutral evolution at synonymous sites will pose problems for attempts to detect selection on proteins, I have not discussed potential solutions. Most notably, can comparing the rates of non-synonymous to (K_a) synonymous (K_s) substitutions still be used to infer the level of selection on proteins? Plotkin et al. (2004) proposed that 'codon volatility', the proportion of mutations which change an amino acid, can be used to detect selection upon a single gene. This method, however, relies on several invalid assumptions (e.g. Chen, Emerson & Martin 2005; Hahn et al. 2005; Nielsen & Hubisz 2005). The general consensus has concluded that the "comparative method rules!" (Dagan & Graur 2005).

As previously described (Chapter 7), detecting selection on proteins using the ratio itself (K_a/K_s) masks the individual contributions of the two rates, with K_a peaks being indistinguishable from K_s dips. Identifying the latter may also allow us to predict functional regions at synonymous sites, as was retrospectively shown in BRCA1 (Orban & Olah 2001). A preliminary mouse-rat comparison showed that, in the 11 out of 143 genes with a significant peak of $K_a/K_s>1$, only half of these peaks could be attributed to positive selection on amino acid changes (L. D. Hurst, unpublished, Chapter 7). It is important to extend this analysis to the scale of whole genomes. Additionally, I can imagine a sliding-window method that asks whether there is a major difference between K_a and K_s while moving along a gene. The significance of the difference would have to be measured relative to the rates of evolution at other sites within the gene.

Conclusion

Evolution at silent sites in mammals has long been thought to be neutral to natural selection. In this thesis, I have shown that this assumption is not valid. Rates of silent site evolution vary between genes, which in part reflect differences in levels of selective constraints. Importantly, synonymous sites do not evolve neutrally. I have provided evidence that synonymous codon usage bias can result from models of selection to optimise mRNA stability and splicing efficiency.

References

Akashi, H., & Eyre-Walker, A. (1998) Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8**: 688-693.

Arndt, P. F., Hwa, T. & Petrov, D. A. (2005) Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J. Mol. Evol.* **60**: 748-763.

Bazykin, G. A., Kondrashov, F. A., Ogurtsov, A. Y., Sunyaev, s. & Kondrashov, A. S. (2004) Positive selection at sites of multiple amino acid replacements since the mouse-rat divergence. *Nature* **429**: 558-562.

Chen, Y., Emerson, J. J. & Martin, T. M. (2005) Evolutionary genomics - Codon volatility does not detect selection. *Nature* **433**: E6-E7.

Clark, A. G. et al. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**: 1960-1963.

Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A. & Stephens, M. (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.* **36**: 700-706.

Dagan, T., & Graur, D. (2005) The comparative method rules! Codon volatility cannot detect positive Darwinian selection using a single genome sequence. *Mol. Biol. Evol.* **22**: 496-500.

Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**: 640-649.

Ellegren, H., Smith, N. G. & Webster, M. T. (2003) Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* **13**: 562-568.

Eskesen, S. T., Eskesen, F. N. & Ruvinsky, A. (2004) Natural selection affects frequencies of AG and GT dinucleotides at the 5' and 3' ends of exons. *Genetics* **167**: 543-550.

Eyre-Walker, A., & Hurst, L. D. (2001) The evolution of isochores. *Nat. Rev. Genet.* **2**: 549-555.

Filipski, J. (1988) Why the rate of silent codon substitution is variable within a vertebrates's genome. *J. Theor. Biol.* **134**: 159-164.

Gaffney, D. J., & Keightley, P. D. (2005) The scale of mutational variation in the murid genome. *Genome Res.*

Hahn, M. W., Mezey, J. G., Begun, D. J., Gillespie, J. H., Kern, A. D., Langley, C. H. & Moyle, L. C. (2005) Evolutionary genomics - Codon bias and selection on single genomes. *Nature* **433**: E5-E6.

Hellmann, I., Ebersberger, I., Ptak, S. E., Paabo, S. & Przeworski, M. (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**: 1527-1535.

Kampa, D. et al. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**: 331-342.

Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. & Gingeras, T. R. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**: 916-919.

Keightley, P. D., Lercher, M. J. & Eyre-Walker, A. (2005) Evidence for Widespread Degradation of Gene Control Regions in Hominid Genomes. *PLoS Biol.* **3**: e42.

Kimura, M., & Ohta, T. (1973) Mutation and evolution at the molecular level. *Genetics* **73**: Suppl 73:19-35.

Kumar, S., & Subramanian, S. (2002) Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99**: 803-808.

Lu, J., Li, W. H. & Wu, C. I. (2003) Comment on "Chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes". *Science* **302**: 988.

Matassi, G., Sharp, P. M. & Gautier, C. (1999) Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**: 786-791.

Mattick, J. S. (2004) RNA regulation: a new genetics? *Nat. Rev. Genet.* **5**: 316-323.

Mattick, J. S., & Makunin, I. V. (2005) Small regulatory RNAs in mammals. *Hum. Mol. Genet.* **14**: R121-132.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. & Donnelly, P. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* **304**: 581-584.

Mikkelsen, T. S. et al. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. **437**: 69-87.

Navarro, A., & Barton, N. H. (2003) Chromosomal speciation and molecular divergence - Accelerated evolution in rearranged chromosomes. *Science* **300**: 321-324.

Nielsen, R. et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**: e170.

Nielsen, R., & Hubisz, M. J. (2005) Evolutionary genomics - Detecting selection needs comparative data. *Nature* **433**: E6-E6.

Okazaki, Y. et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**: 563-573.

Orban, T. I., & Olah, E. (2001) Purifying selection on silent sites - a constraint from splicing regulation? *Trends Genet.* **17**: 252-253.

Perry, J., & Ashworth, A. (1999) Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* **9**: 987-989.

Plotkin, J. B., Dushoff, J. & Fraser, H. B. (2004) Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. *Nature* **428**: 942-945.

Ptak, S. E., Hinds, D. A., Koehler, K., Nickel, B., Patil, N., Ballinger, D. G., Przeworski, M., Frazer, K. A. & Paabo, S. (2005) Fine-scale recombination patterns differ between chimpanzees and humans. *Nat. Genet.* **37**: 429-434.

Semon, M., & Duret, L. (2004) Evidence that functional transcription units cover at least half of the human genome. *Trends Genet.* **20**: 229-232.

Washietl, S., Hofacker, I. L. & Stadler, P. F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA* **102**: 2454-2459.

Willie, E., & Majewski, J. (2004) Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* **20**: 534-538.

Wolfe, K. H., & Li, W. H. (2003) Molecular evolution meets the genomics revolution. *Nat. Genet.* **33**: 255-265.

Zhang, J., Wang, X. & Podlaha, O. (2004) Testing the chromosomal speciation hypothesis for humans and chimpanzees. *Genome Res.* **14**: 845-851.